Enhanced Data Sanitization Model for Effective Privacy in Social Network Data

VIVEKA.R.S.K[1], RAMKUMAR.K[2]
[1]PG Student, Dept of CSE, Bharathiyar College of Engineering and Technology, Karaikal.
[2]Assistant Professor, Dept. of CSE, Bharathiyar College of Engineering and Technology, Karaikal.

Abstract- Social network data can help with obtaining valuable insight into social behaviors and revealing the underlying benefits. New big data technologies are emerging to make it easier to discover meaningful social information from market analysis to counter terrorism. Unfortunately, both diverse social datasets and big data technologies raise stringent privacy concerns. Adversaries can launch inference attacks to predict sensitive latent information, which is unwilling to be published by social users. Therefore, there is a tradeoff between data benefits and privacy concerns. In this paper, we investigate how to optimize the tradeoff between latent-data privacy and customized data utility. As an enhancement from the existing system, we model data sanitization as a game between 1) a publisher who chooses a set of classifiers to apply to data and publishes only instances predicted as non-sensitive and 2) an attacker who combines machine learning and manual inspection to uncover leaked identifying information. We introduce a fast iterative greedy algorithm for the publisher that ensures a low utility for a resource-limited adversary

Keywords- Greedy Algorithm, Social Network Data, Data Mining, Data Sanitization Method, Latent Attributes

1. INTRODCUTION

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

Available@ www.ijrse.org Mail: info@ijrse.org ISSN: 2212-4012

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.

2. RELATED WORK

Preventing Private Information Inference Attacks on Social Networks.

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.

Customized privacy preserving for inherent data and latent data

The huge amount of sensory data collected from mobile devices has offered great potentials to promote more significant services based on user data extracted from sensor readings. However, releasing user data could also seriously threaten user privacy. It is possible to directly collect sensitive information from released user data without user permissions. Furthermore, third party users can also infer sensitive information contained in released data in a latent manner by utilizing data mining techniques. In this paper, we formally define these two types of threats as inherent data privacy and latent data privacy and construct a data-sanitization strategy that can optimize the tradeoff between data utility and customized two types of privacy. The key novel idea lies that the developed strategy can combat against powerful third party users with broad knowledge about users and launching optimal inference attacks. We show that our strategy does not reduce the benefit brought by user data much, while sensitive information can still be protected. To the best of our knowledge, this is the first work that preserves both inherent data privacy and latent data privacy.

Collective Data-Sanitization for Preventing Sensitive Information Inference Attacks in Social Networks

Releasing social network data could seriously breach user privacy. User profile and friendship relations are inherently private. Unfortunately, it is possible to predict sensitive information carried in released data latently by utilizing data mining techniques. Therefore, sanitizing network data prior to release is necessary. In this paper, we explore how to launch an inference attack exploiting social networks with a mixture of non-sensitive attributes and social relationships. We map this issue to a collective classification problem and propose a collective inference model. In our model, an attacker utilizes user profile and social relationships in a collective manner to predict sensitive information of related victims in a released social network dataset. To protect against such attacks, we propose a data sanitization method collectively manipulating user profile and friendship relations. The key novel idea lies that besides sanitizing friendship relations, the proposed method can take advantages of various data-manipulating methods. We show

that we can easily reduce adversary's prediction accuracy on sensitive information, while resulting in less accuracy decrease on non-sensitive information towards three social network datasets. To the best of our knowledge, this is the first work that employs collective methods involving various data-manipulating methods and social relationships to protect against inference attacks in social networks

Customized Privacy Preserving for Classification Based Applications

The rise of sensor-equipped smart phones has enabled a variety of classification based applications that provide personalized services based on user data extracted from sensor readings. However, malicious applications aggressively collect sensitive information from inherent user data without permissions. Furthermore, they can mine sensitive information from user data just in the classification process. These privacy threats raise serious privacy concerns. In this paper, we introduce two new privacy concerns which are inherent-data privacy and latent-data privacy. We propose a framework that enables a data-obfuscation mechanism to be developed easily. It preserves latent-data privacy while guaranteeing satisfactory service quality. The proposed framework preserves privacy against powerful adversaries who have knowledge of users' access pattern and the data-obfuscation mechanism.

Inferring User Demographics and Social Strategies in Mobile Social Networks

Demographics are widely used in marketing to characterize different types of customers. However, in practice, demographic information such as age, gender, and location is usually unavailable due to privacy and other reasons. In this paper, we aim to harness the power of big data to automatically infer users' demographics based on their daily mobile communication patterns.

Our study is based on a real-world large mobile network of more than 7,000,000 users and over 1,000,000,000 communication records (CALL and SMS). We discover several interesting social strategies that mobile users frequently use to maintain their social connections. First, young people are very active in broadening their social circles, while seniors tend to keep close but more stable connections. Second, female users put more attention on cross-generation interactions than male users, though interactions between male and female users are frequent. Third, a persistent same-gender triadic pattern over one's lifetime is discovered for the first time, while more complex opposite-gender triadic patterns are only exhibited among young people.

3. ARCHITECURE

Fig: System Architecture

4.PROPOSED PROCESS

In this work, we explore how to balance the tradeoff between latent-data privacy and data utility. We assume adversaries collect user data, and some privacy-unconscious users publish their sensitive latent information. We first formalize the metrics to measure data utility loss and latent-data privacy. Then, we propose two data sanitization methods that sanitize social attributes and links, respectively. Finally, data-sanitization strategies are proposed, which should not degrade the benefits brought by social

Available@ www.ijrse.org Mail: info@ijrse.org ISSN: 2212-4012

network data, while sensitive latent information can still be protected.

5. MODULES

5.1 Adversary Model:

The powerful adversaries assume with abundant prior knowledge about users, and they can launch optimal inference attacks to infer the SLA of each user. This assumption allows the constructed data-sanitation method can combat against adversaries with a larger range of capability.

First, we assume adversaries know each user's profile. Second, adversaries are assumed to know the data-sanitization strategy employed to realize the tradeoff between utility and privacy. Based on the above knowledge, optimal inference attacks are launched by adversaries.

5.2 Prediction Method for Latent Attributes

The powerful adversaries assume that launch inference attacks by utilizing all publicly available knowledge including social links and attribute sets. Therefore, the prediction method predicts latent information considering social links and attribute sets collectively to increase prediction accuracy.

5.3 Data Sanitization Method

The powerful adversaries are assumed that launch inference attacks by exploiting social links and attribute sets simultaneously. Therefore, in order to realize the tradeoff between privacy and utility, our objective is to sanitize both social links and attribute sets.

1) Attribute-Sanitization Method: An attribute set could be sanitized in three ways, adding attributes, removing attributes, and perturbing attributes (replace one attribute with another). Which methods should be employed to sanitize social data depends on data utility and privacy metrics and data semantics.

2) Link-Sanitization Method: Unlike attributes, social links can only be sanitized by adding links and removing existing links. Similar with the attribute-sanitization method, a link-sanitization method should reduce the prediction accuracy for SLA and do not greatly reduce the prediction accuracy for NSLA. Unfortunately, unlike attributes, it is nontrivial to find the indicative links shared by SLA and NSLA, thus we focus on reducing the prediction accuracy for SLA firstly when sanitize links and more constraints will be given later to guarantee utility.

6. CONCLUSION

In this paper, we study how to optimize the tradeoff between latent-data privacy and customized data utility when combating against powerful adversaries with optimal inference attacks. To address this issue, we first propose two sanitization methods for links and attributes, based on which we formalize prediction utility loss metric, structure utility loss metric and latent-data privacy. Then we formulate an optimization problem that can maximize latent-data privacy while guaranteeing customized data utility. Finally, we evaluate our data-sanitization strategy towards real big social network data and the results show that the proposed data-sanitization strategy can effectively achieve a meaningful privacy-utility tradeoff. Our future work is to explore formal privacy models, such as differential privacy or *k*-anonimity to balance latent-data privacy and customized data utility.

7. REFERENCES

[1] C. Y. Johnson, "Project Gaydar," *Boston Globe*, Sep. 2009.

[2] N. Z. Gong *et al.*, "Joint link prediction and attribute inference using a social-attribute network," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 2, Apr. 2014, Art. no. 27.

[3] Y. Dong, Y. Yang, J. Tang, Y. Yang, and N. V. Chawla, "Inferring user demographics and social strategies in mobile social networks," in *Proc.20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 15–24.

[4] S. Volkova, Y. Bachrach, M. Armstrong, and V. Sharma, "Inferring latent user properties from texts published in social media," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 4296–4297.

[5] A.Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel, "You are who you know: Inferring user profiles in online social networks," in *Proc. 3rd ACM Int. Conf. Web Search Data Mining*, 2010, pp. 251–260.

[6] R. Heatherly, M. Kantarcioglu, and B. M. Thuraisingham, "Preventing private information inference attacks on social networks," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 8, pp. 1849–1862, Aug. 2013.

[7] Z. He, Z. Cai, and Y. Li, "Customized privacy preserving for classification based applications," in *Proc. 1st ACM Workshop Privacy-Aware Mobile Comput.*, 2016, pp. 37–42.

[8] Z. Jorgensen, T. Yu, and G. Cormode, "Publishing attributed social graphs with formal privacy guarantees," in *Proc. Int. Conf. Manage. Data*, 2016, pp. 107–122.

[9] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao, "Privbayes: Private data release via Bayesian networks," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2014, pp. 1423–1434.

[10] C. Liu and P. Mittal, "Linkmirage: Enabling privacy-preserving analytics on social relationships," in *Proc. Netw. Distrib. Syst. Security Symp.*, 2016, pp. 492–503.

[11] W.-Y. Day, N. Li, and M. Lyu, "Publishing graph degree distribution with node differential privacy," in *Proc. Int. Conf. Manage. Data*, 2016, pp. 123–138.

[12] P. Gundecha, G. Barbier, J. Tang, and H. Liu, "User vulnerability and its reduction on a social networking site," *ACM Trans. Knowl. Discovery Data*, vol. 9, no. 2, Sep. 2014, Art. no. 12.

[13] M. Han, M. Yan, Z. Cai, and Y. Li, "An exploration of broader influence maximization in timeliness networks with opportunistic selection," *J. Netw. Comput. Appl.*, vol. 63, pp. 39–49, 2016.

[14] Z. He, Z. Cai, Q. Han, W. Tong, L. Sun, and Y. Li, "An energy efficient privacy-preserving content sharing scheme in mobile social networks," *Pers. Ubiquitous Comput.*, vol. 20, no. 5, pp. 833–846, 2016.

[15] Z. He, Z.Cai,Y. Sun,Y. Li, and X. Cheng, "Customized privacy preserving for inherent data and latent data," *Pers. Ubiquitous Comput.*, vol. 21, no. 1, pp. 43–54, Feb. 2017.

[16] Z. Cai, Z. He, X. Guan, and Y. Li, "Collective data-sanitization for preventing sensitive information inference attacks in social networks," *IEEE Trans. Dependable Secure Comput.*, to be published.

[17] Z. He, Y. Li, J. Li, J. Yu, H. Gao, and J. Wang, "Addressing the threats of inference attacks on traits and genotypes from individual genomic data," in *Proc. Int. Symp. Bioinformat. Res. Appl.*, 2017, pp. 223–233.

[18] X. Zheng, Z. Cai, J. Li, and H. Gao, "A study on application-aware scheduling in wireless networks," *IEEE Trans. Mobile Comput.*, vol. 16, no. 7, pp. 1787–1801, Jul. 2017.

[19] X. Zheng and Z. Cai, "Real-time big data delivery in wireless networks: A case study on video delivery," *IEEE Trans. Ind. Informat.*, vol. 13, no. 4, pp. 2048–2057, Aug. 2017.

[20] X. Zheng, Z. Cai, J. L, and H. G, "Location-privacy-aware review publication mechanism for local business service systems," in *Proc. 36th Annu. IEEE Int. Conf. Comput. Commun.*, 2017.

[21] T. Qiu, R. Qiao, and D. Wu, "EABS: An event-aware backpressure scheduling scheme for emergency internet of things," *IEEE Trans. Mobile Comput.*, to be published.

[22] T. Qiu, A. Zhao, R.Ma, V. Chang, F. Liu, and Z. Fu, "A task-efficient sink node based on embedded multi-core SOC for internet of things," *Future Gener. Comput. Syst.*, 2016.

[23] Y. E. Sun *et al.*, "Privacy-preserving strategy proof auction mechanisms for resource allocation," *Tsinghua Sci. Technol.*, vol. 22, no. 2, pp. 119–134, Apr. 2017.

[24] J. Lu and X. Wang, "Interference-aware probabilistic routing for wireless sensor networks," *Tsinghua Sci. Technol.*, vol. 17, no. 5, pp. 575–585, 2012.

[25] Z. He, Y. Li, and J. Wang, *Differential Privacy Preserving Genomic Data Releasing via Factor Graph*. New York, NY, USA: Springer, 2017, pp. 350–355.

[26] Z. He, Z. Cai, J. Yu, X.Wang, Y. Sun, and Y. Li, "Cost-efficient strategies for restraining rumor spreading in mobile social networks," *IEEE Trans.Veh. Technol.*, vol. 66, no. 3, pp. 2789–2800, Mar. 2017.

[27] C. Dwork, "Differential privacy," in *Automata, Languages and Programming (ser. Lecture Notes in Computer Science)*, vol. 4052, M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, Eds. Berlin, Germany: Springer, 2006, pp. 1–12.

[28] L. Sweeney, "K-anonymity:Amodel for protecting privacy," *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 557–570, Oct. 2002.

[29] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "L-diversity: Privacy beyond k-anonymity," *ACM Trans. Knowl. Discovery Data*, vol. 1, no. 1, Mar. 2007, Art. no. 3.

[30] M. Sviridenko, "A note on maximizing a submodular set function subject to a knapsack constraint," *Oper. Res. Lett.*, vol. 32, no. 1, pp. 41–43, 2004.