A REINFORCED CUSTOMIZED PRIVACY PROTECTION RECOMMENDATIONMODEL WITH USER TRUST

BHUVANA.A[1], DINESH KUMAR.R[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*—** Personalized recommendation is crucial to help users find pertinent information. It often relies on a large collection of user data, in particular users' online activity (e.g., tagging/rating/checking-in) on social media, to mine user preference. However, releasing such user activity data makes users vulnerable to inference attacks, as private data (e.g., gender) can often be inferred from the users' activity data. In this paper, we proposed PrivRank, a customizable and continuous privacy-preserving social media data publishing framework protecting users against inference attacks while enabling personalized ranking-based recommendations. Its key idea is to continuously obfuscate user activity data such that the privacy leakage of user-specified private data is minimized under a given data distortion budget, which bounds the ranking loss incurred from the data obfuscation process in order to preserve the utility of the data for enabling recommendations. An empirical evaluation on both synthetic and real-world datasets shows that our framework can efficiently provide effective and continuous protection of user-specified private data, while still preserving the utility of the obfuscated data for personalized ranking-based recommendation. Compared to state-of-the-art approaches, PrivRank achieves both a better privacy protection and a higher utility in all the ranking-based recommendation use cases we tested**.**

Keywords- PrivRank, PrivCheck, Nearest neighbor method, Ranking-Based Recommendation, Historical Data Publishing, Online Data Stream Publishing

1.INTRODCUTION

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.

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

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

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. 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. 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).

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. 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.

2 RELATED WORK

PrivCheck: privacy-preserving check-in data publishing for personalized location based services.

With the widespread adoption of Smartphone's, we have observed an increasing popularity of Location-Based Services (LBSs) in the past decade. To improve user experience, LBSs often provide personalized recommendations to users by mining their activity (i.e., check-in) data from location-based social networks[2]. However, releasing user check-in data makes users vulnerable to inference attacks, as private data (e.g., gender) can often be inferred from the users' check-in data. In this paper, we propose PrivCheck, a customizable and continuous privacy-preserving check-in data publishing framework providing users with continuous privacy protection against inference attacks. The key idea of PrivCheck is to obfuscate user check-in data such that the privacy leakage of user-specified private data is minimized under a given data distortion budget, which ensures the utility of the obfuscated data to empower personalized LBSs. Since users often give LBS providers

access to both their historical check-in data and future check-in streams, we develop two data obfuscation methods for historical and online check-in publishing, respectively. An empirical evaluation on two real-world datasets shows that our framework can efficiently provide effective and continuous protection of user-specified private data, while still preserving the utility of the obfuscated data for personalized LBSs.

Protecting Individual Information Against Inference Attacks in Data Publishing

In many data-publishing applications, the data owner needs to protect sensitive information pertaining to individuals. Meanwhile, certain information is required to be published. The sensitive information could be considered as leaked, if an adversary can infer the real value of a sensitive entry with a high confidence. In this paper we study how to protect sensitive data when an adversary can do inference attacks using association rules derived from the data[3]. We formulate the inference attack model, and develop complexity results on computing a safe partial table. We classify the general problem into subcases based on the requirements of publishing information, and propose the corresponding algorithms for finding a safe partial table to publish. We have conducted an empirical study to evaluate these algorithms on real data.

Performance of recommender algorithms on top-N recommendation tasks

In many commercial systems, the 'best bet' recommendations are shown, but the predicted rating values are not. This is usually referred to as a top-N recommendation task, where the goal of the recommender system is to find a few specific items which are supposed to be most appealing to the user. [6] Common methodologies based on error metrics (such as RMSE) are not a natural fit for evaluating the top-N recommendation task. Rather, top-N performance can be directly measured by alternative methodologies based on accuracy metrics (such as precision/recall). An extensive evaluation of several state-of-the art recommender algorithms suggests that algorithms optimized for minimizing RMSE do not necessarily perform as expected in terms of top-N recommendation task. Results show that improvements in RMSE often do not translate into accuracy improvements.

In particular, a naive non-personalized algorithm can outperform some common recommendation approaches and almost match the accuracy of sophisticated algorithms. Another finding is that the very few top popular items can skew the top-N performance. The analysis points out that when evaluating a recommender algorithm on the top-N recommendation task, the test set should be chosen carefully in order to not bias accuracy metrics towards non-personalized solutions. Finally, we offer practitioners new variants of two collaborative filtering algorithms that, regardless of their RMSE, significantly outperform other recommender algorithms in pursuing the top-N recommendation task, with offering additional practical advantages. This comes at surprise given the simplicity of these two methods.

Managing your Private and Public Data: Bringing down Inference Attacks against your Privacy

The practical methodology is proposed to protect a user's private data, when he wishes to publicly release data that is correlated with his private data, to get some utility. Our approach relies on a general statistical inference framework that captures the privacy threat under inference attacks, given utility constraints. Under this framework[15], data is distorted before it is released, according to a probabilistic privacy mapping. This mapping is obtained by solving a convex optimization problem, which minimizes information leakage under a distortion constraint. We address practical challenges encountered when applying this

theoretical framework to real world data. On one hand, the design of optimal privacy mappings requires knowledge of the prior distribution linking private data and data to be released, which is often unavailable in practice. On the other hand, the optimization may become untraceable when data assumes values in large size alphabets, or is high dimensional. Our work makes three major contributions.

First, we provide bounds on the impact of a mismatched prior on the privacy-utility tradeoff. Second, we show how to reduce the optimization size by introducing a quantization step, and how to generate privacy mappings under quantization. Third, we evaluate our method on two datasets, including a new dataset that we collected, showing correlations between political convictions and TV viewing habits. We demonstrate that good privacy properties can be achieved with limited distortion so as not to undermine the original purpose of the publicly released data, e.g., recommendations.

Ranking Measures and Loss Functions in Learning to Rank

Learning to rank has become an important research topic in machine learning. While most learning-to-rank methods learn the ranking functions by minimizing loss functions, it is the ranking measures (such as NDCG and MAP) that are used to evaluate the performance of the learned ranking functions. In this work, we reveal the relationship between ranking measures and loss functions in learning- to-rank methods, such as Ranking SVM, RankBoost, RankNet, and ListMLE[16]. We show that the loss functions of these methods are upper bounds of the measure- based ranking errors. As a result, the minimization of these loss functions will lead to the maximization of the ranking measures. The key to obtaining this result is to model ranking as a sequence of classification tasks, and define a so-called essential loss for ranking as the weighted sum of the classification errors o f individual tasks in the sequence. We have proved that the essential loss is both an upper bound of the measure-based ranking errors, and a lower bound of the loss functions in the aforementioned methods. Our proof technique also suggests a way to modify existing loss functions to make them tighter bounds of the measure-based ranking errors. Experimental results on benchmark datasets show that the modifications can lead to better ranking performances, demonstrating the correctness of our theoretical analysis.

3.PROPOSED SYSTEM

In this paper, the problem of privacy-preserving publishing of user social media data by considering both the specific requirements of user privacy on social media and the data utility for enabling high-quality personalized recommendation is studied .

PrivRank a new framework, a customizable and continuous privacy-preserving data publishing framework is proposed to protect users against inference attacks while enabling personalized ranking based recommendation.

First, considering the use case of recommendation based on social media data, we identify a privacy-preserving data publishing problem by analyzing the specific privacy requirements and users' benefits of social media.

Second, a customizable and continuous data obfuscation framework is proposed for user activity data on social media. The key idea is to measure the privacy leakage of user-specified private data from public data based on mutual information, and then to obfuscate public data such that the privacy leakage is minimized under a given data distortion budget, which can ensure the utility of the released data.

Third, to guarantee the utility of the obfuscated data for enabling personalized

ranking-based recommendation, we measure and bound the data distortion using a pair wise ranking loss metric, i.e., the Kendall-_ rank distance. To efficiently incorporate such ranking loss, we propose a bootstrap sampling process to fast approximate the Kendall-_ distance.

It provides continuous protection of user-specified private data against inference attacks by obfuscating both the historical and streaming user activity data before releasing them, while still preserving the utility of the published data for enabling personalized ranking based recommendation by efficiently limiting the pair wise ranking loss incurred from data obfuscation.

The results show that PrivRank can continuously provide customized protection of user-specified private data, while the obfuscated data can still be exploited to enable high-quality personalized ranking based recommendation

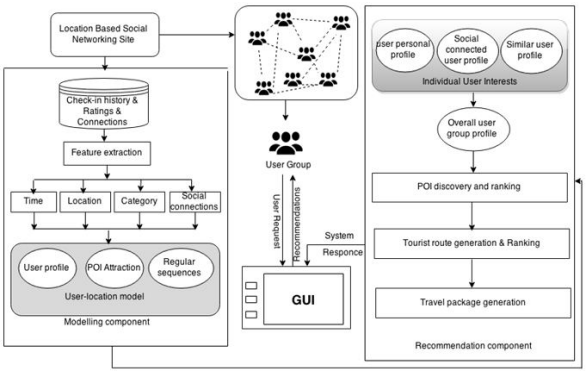

Fig 1.System Architecture

3.1 User Preference Modeling

Users' activities on social media massively imply their preferences. Individual social media services often provide users with a unique feature (or a certain type of items) for interaction, such as photos for Flickr, videos for YouTube, music for Last.fm, and POIs for Foursquare. By interacting with these items on social media (e.g., tagging a photo, rating a video or checking-in at a POI), users explicitly or implicitly express their preferences on those items.

3.2 Ranking-Based Recommendation

Ranking based recommendation outputs a ranked list of items for a user, where the top items are most likely to be appealing to her. The related algorithms mainly leverage the existing ranking of items in the learning process to predict the missing rank of the items for recommendation. Therefore, ranking-based recommendation algorithms are sensitive to the ranking loss incurred from the data obfuscation process, rather than other types of loss measured by the Euclidean or Squared L2 distance, for example. Moreover, those traditional data distortion measures are not analogous to ranking loss. Therefore, considering ranking loss incurred from data obfuscation is critical for ranking based recommendation.

3.3 Historical Data Publishing

To publish historical public data in a privacy-preserving way, the key idea is to probabilistically obfuscate a user's historical public data vector to that of another user, which are similar but have less privacy leakage. In this context, data obfuscation operates on one's whole public data vector, rather than obfuscating her individual activity records one by one (over the user's activity stream). Compared to the streaming scheme, we show that such a historical data obfuscation scheme can achieve the same level of privacy protection with a lower data distortion budget.

3.4 Online Data Stream Publishing

The service providers have access to the user's future activity streams. Therefore, we protect her private data by obfuscating her activity stream on-the-fly. Different from historical data publishing, the streaming nature of user activity imposes the following constraint on online data obfuscation: Due to time and space efficiency requirements of real-time data publishing (i.e., single-pass processing with limited memory), online

data obfuscation can only be performed based on the incoming activity data instance itself (e.g., a new rating/tagging/checking-in activity on an item), without accessing the user's historical data. In other words, we want to obfuscate each activity to another with less privacy leakage. However, a rating/tagging/checking-in activity of a user will probably lead to a certain modification of the user's public data vector that encodes a certain type of user preference, such as the user's ratings (on a 1-5 scale), her tags/thumbs-ups (in a binary format), or her cumulative number of interactions (e.g., the number of check-ins on POIs).

Conclusion and Future Enhancement

This paper introduced PrivRank, a customizable and continuous privacy-preserving social media data publishing framework. It continuously protects user-specified data against inference attacks by releasing obfuscated user activity data, while still ensuring the utility of the released data to power personalized ranking-based recommendations. To provide customized protection, the optimal data obfuscation is learned such that the privacy leakage of user-specified private data is minimized; to provide continuous privacy protection, we consider both the historical and online activity data publishing; to ensure the data utility for enabling ranking-based recommendation, we bound the ranking loss incurred from the data obfuscation process using the Kendall-_ rank distance. We showed through extensive experiments that PrivRank can provide an efficient and effective protection of private data, while still preserving the utility of the published data for different ranking-based recommendation use cases

REFERENCES

[1] S. Salamatian, A. Zhang, F. du Pin Calmon, S. Bhamidipati, N. Fawaz, B. Kveton, P. Oliveira, and N. Taft, "How to hide the elephant-or the donkey-in the room: Practical privacy against statistical inference for large data," in Proc. of GlobalSIP. IEEE, 2013.

[2] D. Yang, D. Zhang, Q. Bingqing, and P. Cudre-Mauroux, "Privcheck: Privacy-preserving check-in data publishing for personalized location based services," in Proc. of UbiComp'16. ACM, 2016.

[3] C. Li, H. Shirani-Mehr, and X. Yang, "Protecting individual information against inference attacks in data publishing," in Advances in Databases: Concepts, Systems and Applications. Springer, 2007, pp. 422–433.

[4] B. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," ACM Computer Survey, vol. 42, no. 4, p. 14, 2010.

[5] I. A. Junglas, N. A. Johnson, and C. Spitzm ¨ uller, "Personality traits and concern for privacy: an empirical study in the context of location-based services," European Journal of Information Systems, vol. 17, no. 4, pp. 387–402, 2008.

[6] P. Cremonesi, Y. Koren, and R. Turrin, "Performance of recommender algorithms on top-n recommendation tasks," in Proc. Of RecSys'10. ACM, 2010, pp. 39–46.

[7] N. Li, R. Jin, and Z.-H. Zhou, "Top rank optimization in linear time," in Advances in neural information processing systems, 2014, pp. 1502–1510.

[8] M. G. Kendall, "Rank correlation methods." 1948.

[9] L. Sweeney, "k-anonymity: A model for protecting privacy," International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 05, pp. 557–570, 2002.

[10] L. Sankar, S. R. Rajagopalan, and H. V. Poor, "Utility-privacy tradeoffs in databases:

An information-theoretic approach," IEEE Transactions on Information Forensics and Security, vol. 8, no. 6, pp. 838–852, 2013.

[11] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," ACM Transactions on Knowledge Discovery from Data, vol. 1, no. 1, p. 3, 2007.

[12] C. Dwork, "Differential privacy," in Automata, languages and programming. Springer, 2006, pp. 1–12.

[13] F. du Pin Calmon and N. Fawaz, "Privacy against statistical inference," in Proc. of Allerton'12. IEEE, 2012, pp. 1401–1408.

[14] A. Zhang, S. Bhamidipati, N. Fawaz, and B. Kveton, "Priview: Media consumption and recommendation meet privacy against inference attacks," IEEE Web, vol. 2, 2014.

[15] S. Salamatian, A. Zhang, F. du Pin Calmon, S. Bhamidipati, N. Fawaz, B. Kveton, P. Oliveira, and N. Taft, "Managing your private and public data: Bringing down inference attacks against your privacy," IEEE Journal of Selected Topics in Signal Processing, vol. 9, no. 7, pp. 1240–1255, 2015.

[16] W. Chen, T.-Y. Liu, Y. Lan, Z.-M. Ma, and H. Li, "Ranking measures and loss functions in learning to rank," in Proc. of NIPS, 2009, pp. 315–323.

[17] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," PNAS, vol. 95, no. 25, pp. 14 863–14 868, 1998.

[18] R. Baeza-Yates, B. Ribeiro-Neto et al., Modern information retrieval. ACM press New York, 1999, vol. 463.

[19] K. J¨arvelin and J. Kek¨al¨ainen, "Cumulated gain-based evaluation of ir techniques," ACM Transactions on Information Systems (TOIS), vol. 20, no. 4, pp. 422–446, 2002.

[20] B. Efron and R. J. Tibshirani, An introduction to the bootstrap. CRC press, 1994.

[21] M. Grant and S. Boyd, "Graph implementations for nonsmooth convex programs," in Recent Advances in Learning and Control. Springer, 2008, pp. 95–110.

[22] G. S. Manku, S. Rajagopalan, and B. G. Lindsay, "Approximate medians and other quantiles in one pass and with limited memory," in ACM SIGMOD Record, vol. 27, no. 2. ACM, 1998, pp. 426–435.

[23] D. Yang, D. Zhang, Z. Yu, and Z. Wang, "A sentiment-enhanced personalized location recommendation system," in Proc. of HT'13. ACM, 2013, pp. 119–128.

[24] D. Yang, D. Zhang, V. W. Zheng, and Z. Yu, "Modeling user activity preference by leveraging user spatial temporal characteristics in lbsns," IEEE Transactions on System, Man, Cybernetics: System, vol. 45, no. 1, pp. 129–142, 2015.

[25] Z. Yu, H. Xu, Z. Yang, and B. Guo, "Personalized travel package with multi-point-of-interest recommendation based on crowdsourced user footprints," IEEE Transactions on Human-Machine Systems, vol. 46, no. 1, pp. 151–158, 2016.

[26] D. Yang, D. Zhang, L. Chen, and B. Qu, "Nationtelescope: Monitoring and visualizing large-scale collective behavior in lbsns," Journal of Network and Computer Applications, vol. 55, pp. 170–180, 2015.

[27] D. Yang, D. Zhang, and B. Qu, "Participatory cultural mapping based on collective behavior data in location-based social networks," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 7, no. 3, p. 30, 2016.

[28] Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui, "Exploring millions of footprints in location sharing services." Proc. of ICWSM'11, vol. 2011, pp. 81–88, 2011.

[29] X. Zhao, L. Li, and G. Xue, "Checking in without worries: Location privacy in location based social networks," in Proc. of INFOCOM' 13. IEEE, 2013, pp. 3003–3011.

[30] C. X. Ling, J. Huang, and H. Zhang, "Auc: a better measure than accuracy in comparing learning algorithms," in Advances in Artificial Intelligence. Springer, 2003, pp. 329–341.

[31] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "Bpr: Bayesian personalized ranking from implicit feedback," in Proc. of UAI'09. AUAI Press, 2009, pp. 452–461.

[32] D. Yang, D. Zhang, Z. Yu, and Z. Yu, "Fine-grained preference aware location search leveraging crowdsourced digital footprints from lbsns," in Proc. of UbiComp'13. ACM, 2013, pp. 479–488.

[33] S. Agrawal and J. R. Haritsa, "A framework for high-accuracy privacy-preserving mining," in Proc. of the ICDE'05. IEEE, 2005, pp. 193–204.

[34] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in Theory of Cryptography Conference. Springer, 2006, pp. 265–284.

[35] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in Proc. of FOCS'07. IEEE, 2007, pp. 94–103.

[36] R. Chen, N. Mohammed, B. C. Fung, B. C. Desai, and L. Xiong, "Publishing set-valued data via differential privacy," PVLDB, vol. 4, no. 11, pp. 1087–1098, 2011.

[37] L. Wang, D. Yang, X. Han, T. Wang, D. Zhang, and X. Ma, "Location privacy-preserving task allocation for mobile crowdsensing with differential geo-obfuscation," in Proceedings of the 26th International Conference on World Wide Web. ACM, 2017, pp. 627–636.

[38] C. Dwork, "Differential privacy: A survey of results," in Proc. Of TAMC. Springer, 2008, pp. 1–19.

[39] Z. Huang and S. Kannan, "The exponential mechanism for social welfare: Private, truthful, and nearly optimal," in Proc. of FOCS'12, 2012.

[40] Y. Shen and H. Jin, "Privacy-preserving personalized recommendation: An instance-based approach via differential privacy," in Proc. of ICDM. IEEE, 2014, pp. 540–549.