A Review of Existing Approaches Affecting Privacy in Social Networking

Dr Ratandeep Kaur

Assistant Professor, IT Department
SGTBIMIT, Delhi, India

Abstract : Social networking has become a largest platform for data mining in today 's environment to provide the privacy to users. Online social networks allow their users to share the information related to their personal lives, to communicate in various ways and upload multimedia content over the public platform. Information can easily be disclosed to an unintended wider audience due to this kind of exposure. The unlimited access to the information over Internet along with another online application has introduced a novel area of research where data mining algorithms must consider with the perspective of privacy preservation, called privacy-preserving data mining (PPDM).A brief review of literature has been discussed in this paper related to privacy preserving data mining along with social network.

Keywords: PPDM, Social networking, Anonymization

I. INTRODUCTION

In recent years, social networking has become the growing trend for youth as well as adult age groups. Most of these users often check and change their privacy settings for their public profiles. A user profile includes the information with the intention of an individual to share over a social network site. In most scenarios, users have preferences to share location, address, and e-mail and phone numbers on their profiles. Users can also post information regarding their education, employment, personal interests, and other insignificant information, like favourite movies and music **(Timm, 2008)**. In present environment, social networking organization wants to improve their existing system and take support of machine learning algorithms in their decision-making process. Users 'privacy prediction helps social media providers to take corrective actions for the users having a higher level of privacy risk so that these users can share their limited set of information on the public platform. Therefore, the research questions that arise pertaining to users 'privacy analysis can be as follows:

- What are the relationships among the attributes for users 'privacy level prediction to obtain higher privacy?
- How a probabilistic model for the users'privacy level prediction is generated for applying the classification techniques?
- How can the knowledge obtain from profile data of user assist decision makers to progress decision-making processes?

II.BACKGROUND

PPDM has become prominent as it enables the mining of bulk of data related to users while protecting their sensitive information. As the purpose of data mining is to acquire valuable information or gaining knowledge from numerous data source, whereas the privacy-preservation in data mining is apply to preserve data beside any kind of leak or information loss. PPDM is a broad area to explore within data mining as new challenges are emerging as a result of the increase in usage of social network sites. **Wang *et al.* (2009**) wrote PPDM is mainly focused on reducing the privacy risk while amending the data in a way that sensitive information can be protected when performing data mining operations. Privacy preserving data mining is a dual approach. First, sensitive data like users' ID, name, contact, address, etc., ought to be modifies or removes from the original database, so that the receiver of that information cannot compromise the privacy of the user. Secondly, sensitive knowledge gained from data mining algorithm applied to the database should also be excluded, as it can also compromise the privacy of user's data (**Verykios *et al.*, 2004).**

Existing PPDM Approaches

In recent years, various PPDM techniques have been proposed and developed as to protect sensitive data of the data owners. The main role of privacy preserving data mining is to sustain stability among information loss along with privacy loss in order to protect sensitive data from disclosure whereas at the same time maintains the accuracy of data mining results. Nonetheless, not a solitary strategy is available that is reliable in all spaces and can resolve some unsolved problems in the future. As the majority of the algorithms were developed for centralized database. However, in today's scenario, the expansion of digital environment results in data storage at different sites in a distributed database.A lot of algorithms focus on protecting the individual's private information, however, does not concentrate on the security of sensitive information. There is no particular technique which can attain both data hiding along with rule hiding. Every algorithm works on performing single functionality of data mining. There is yet no single method available that can perform all type of data mining task **(Shelke *et al.*, 2015)**. The solution to overcome the limitations of the various PPDM techniques two or more techniques are merged. This new approach is introduced as hybrid technique in which many algorithms have been proposed to merge multiple techniques.

Kantarcioglu *et al.* (2004) proposed a hybrid method to connect noise addition and SMC for gradually perform association mining over horizontally partitioned data. In this method, while multiple parties have shared set of encrypted keys, added a slightest noise in the data to alter it as false key. These approaches become a sustainable solution for preserving privacy by transforming the original data, but also result in excessive loss of information. In future, there is a scope for effective tools and techniques which incorporate anonymization of various sensitive attributes, evaluation of huge datasets along with non-homogeneous data anonymization for attainment of minimum information loss and accuracy of released data.Therefore, the concept of Data Mining comes into light as an interdisciplinary field consisting of statistical measure and machine learning techniques for supporting data

analysis and prediction. Data mining is an inductive approach of analyzing data where machine learning algorithms are employed to acquire knowledge from data can also be used to design and develop algorithms for privacy preservation.

Privacy over Social Network

Social networking sites gaining popularity as they are providing the medium of communication between different users along with the full range of services to users such as to share their personal as well as social information like their whereabouts, photos, relationships, messaging and so on **(Gross *et al.*, 2005).** These sites are helping people to maintain their relationship with others and give them a platform to share their views, feelings, like or dislikes related to their social circle.

Social networking sites are providing the interface which can be easily understandable and make it easy to use which results in increasing number of users. The day-to-day rising number of users of social networking sites is leading to issues related to the privacy and security of their users **(Fogel *et al.*, 2009)**. In recent years social networks such as Facebook, Instagram, Twitter, WhatsApp, etc., have become more popular among the various sites which are in use frequently by the people all around the world **(Dutton, 2004)**. In the past years, data mining techniques had developed for delivering privacy preservation. PPDM is gaining more popularity as it enables the sharing of data related to private and sensitive information of a user **(Provost *et al.*, 2009)**.

III.LITERATURE REVIEW

This section covers the work done so far in the privacy preservation in social networking sites. The Literature Review is presented in three subsections including Privacy issues in social network, followed by PPDM in Social networking and ends with most popular current techniques namely extended or hybrid approaches for pertaining privacy to social networking sites.

Privacy issues in Social network

Dolvara Gunatilaka (2011) wrote, "most of the users of online social network applications use their real name as the profile name. Thus, users' name is openly accessible on social media and all the social media are recorded in the web indexes. Unauthorized users can get all the potential data of the individual through social networking websites. Social network's users nearly give their actual name and sensitive information on their profile, for example, name, contact information, date of birth, relationship status, education details, present and previous work locations."**Weiss (2008)** discussed the comparison between traditional and new approach of privacy demands by online social networks. In the customary Web, privacy is kept up through excessive data collection, hiding individual's identity and hardly allowing access to approved parties. The fact of OSNs is that information and identity are firmly connected and normally noticeable to large groups of individuals. By way of further information getting accessible online it is diligently for an individual to monitor and control that information. **Rosenblum (2007)** cited data in social networking sites is enormously available to a extensive viewers than apparent to its owners and land up in the media. **Gross *et al.* (2005)** cited that majority of the social network users don't change the default settings for privacy as given by those sites and furthermore share a lot of data on their profile that can prompt privacy outflow.

Palen *et al.* (2003) classified various privacy issues faced by individuals. First, the *disclosure issues* used for controlling the pressure amid private and public. Second, *identity issues* for overseeing self-portrayal in front of a explicit viewers. Third, the *temporal issues* which handle previous behaviour with future opportunities. **Kang (1998)** defined privacy as a person has rights to control the circumstances under which individual data i.e. data recognizable to the individual is obtained, unveiled, or utilized.

Privacy Preserving Data mining

Aghasian *et al.* (2018) proposed a new privacy-preservation approach through accepting friend request method that enables individual to be sure about data to share with others while having low risk of being oppressed or identified. **Ganesan *et al.* (2017)** stated that social networking sites increase the rate of cybercrime and proposed k-means clustering approach to analyze the data. **Kumar *et al.* (2016)** surveyed the data available on various domains, concluded that undisclosed information of user should be private.

Niu *et al.* (2015) proposed, "A new attack named Variance Based Attack (VBA) on short-range communication-based spatial cloaking algorithms in order to preserve privacy in location-based services supported by the Internet of things." **Jiang *et al.* (2015)** proposed a safe and extensible storage mechanism with concern of security, scalability, flexibility and reliability to satisfy the requirements for data mining and analytics with large aggregate data. **Xi *et al.* (2015)** proposed a methodology of secure administration structure alongside data flow control for accomplishing basic services with the goal of diminishing complexity and cost of verification. **Li *et al.* (2014)** conducted a study relied upon a PPDM framework while comparing the benefits and drawbacks of various PPDM technique in social networking.**Cheng *et al.* (2013)** formed an outline to give client's control in such a way through which third-party applications can get to their information as well as activities in social networks however as yet keeping the usefulness of third-party applications to preserve user's privacy.

Heatherly *et al.* (2013) conducted a study based on a classification algorithm known as Naive Bayes that utilizes node descriptions and their relation with each other, to predict private parameters, analyse user's profile as well as their friendship links along with their other details to provide better predictability than details alone. Their research concluded that for protection of privacy then both user's details and link details must be sanitized. It means, eliminating some data from a user's profile and removing links across friends.**Beck *et al.* (2012)** developed a software means to sustain the anonymizing method of PPDM to compose a demonstrator through a user-friendly interface and achieves anonymization to enhance the utility by swapping and recording. **Yang (2011)** proposed that non-critical and comprehensive information has been used to sustain social network analysis and mining to provide privacy protection of information. **Wondracek *et al.* (2010)** described method by introducing a novel, practical de-anonymization attack that makes use of the group information in social networking sites and concluded that by using information about group members and access the history, attackers could reveal anonymity of social network users. **Zhu *et al.* (2010)** proposed "a combined framework for controlling access in social networks through pioneering key management." **Ding *et al.* (2010)** presented an analysis of the attacks on de-anonymization which occurs in social networks.

Lan *et al.* (2010) conducted a study on synthetic dataset to propose a technique to preserve privacy of social network's users through graphs. **Tootoonchian *et al.* (2009)** proposed, "A software tool that improves the privacy of centralized and decentralized online

systems such as content sharing networks like BitTorrent." Research has been conducted by **(Lijie *et al.*, 2009)** to study "the relationship identification in which the more vulnerable attacks was taken into account using link probability." **Ford *et al.* (2009)** employed "p-sensitive k-anonymity algorithm" to analyze social network based on a greedy-clustering approach. Research has been conducted by **(Lin *et al.*, 2009)**, to identify the problem of privacy-preserving mining of numerous item sets. They proposed an algorithm based on association rule to protect the data through addition of noisy data to each transaction and got significant high accuracy.**Zhou *et al.* (2008)** conducted a study on PPDM in the context of social networking and identified that estimating the loss of information while anonymizing social network data is complicated than anonymizing relational data.**Boyd (2004)** described social networks as web applications that facilitate their clients to form their semi-public profile, i.e. a profile where some information is public and some is private, communicate with those who are their friends and made an online community. It is based on social relationships among users. **Erkin *et al.* (2007)** conducted a study on image and signal processing where the problem of security is vigorously caused by using k -means clustering approaches and concluded that to preserve privacy in the k-means algorithm, there must be a secure multi-party relation that establish a formal model to protect privacy of data.

Extend approach of Privacy preserving data mining

Nergiz *et al.* (2013) introduced the hybrid generalizations. It is not only performed the generalizations but also implicated the mechanism for data transfer. In the data process, changed the position of certain cells to some populated indistinguishable data cells. The relocation process helped to generate anonymization of finer granularity and ensured underlying privacy. The data relocation is a trade-off among the utilization and reliability of the data, where controlled the trade-off by the provider parameter. The results revealed that a small number of relocations could enhance the utility as compared to the heuristic metrics and query answering accuracy.**Zhang *et al.* (2013)** developed a hybrid approach along with Top-Down Specialization (TDS) and Bottom-Up Generalization (BUG) techniques. In this method, one of the two components is selected automatically by comparing-anonymity parameter with workload balancing point which is defined by the clients. Both TDS and BUG are obtained in a scalable way via a series of deliberately designed Map Reduce jobs. Based on the contributions herein, it is worth exploring the next step on scalable privacy preservation aware analysis and scheduling on large-scale datasets. **Lohiya *et al.* (2012)** proposed a hybrid technique in which they used randomization and generalization. In this approach, data is randomized and then modifies. This technique protects data with better accuracy as well as it can restructure original data and supply data with no information loss. **Kavianpour *et al.* (2011)** "designed an integrated algorithm by consolidating the benefits of k-anonymity and l-diversity algorithm at that point assessed the adequacy of the joined qualities. This algorithm has option to build the dimension of privacy for social network users by anonymizing and diversifying revealed data. **Tang *et al.* (2010)** used algorithm of data mining for building generalized sub graphs prior to sharing the social network with other parties and a method to incorporate the generalized data to discover the closeness centrality measures."

IV.CHALLENGES IN EXISTING APPROACHES FOR PERTAINING TO PRIVACY IN SOCIAL NETWORKING

Data mining techniques have been used broadly in both centralized and distributed data environments. Though, it is widely known that data mining may cause a threat to security and particularly privacy yet it may be likely to disclose sensitive information of individuals. For example, in a distributed data environment, data mining may enable involving parties to reveal each other's sensitive information that was not intended to be shared. **Zheleva *et al.* (2007)** proposed, "an audience segregation model based on social interaction and derived that allowing strangers to join user's friend list can lead to privacy risk. However, this model can only support single binary relational ties(e.g., friends or stranger) but human relationships are more complex required grouping mechanism." In **(Backstrom *et al.*, 2007)** research, an attack was taken into account against the anonymized network. As the network consists of only nodes and edges, they ignore detail values that can identify people.

Hay *et al.* (2007) and **Liu *et al.* (2008)** consider various ways of anonymizing social networks but focuses on inferring details from nodes in the network, not individually identifying individuals. **Gkoulalas *et al.* (2009)** introduced a novel approach to anonymize the data to minimize the information loss by satisfying the utilization of data publisher. **Oliveira *et al.* (2010)** proposed a technique to randomised data by adding noise with a known statistical distribution. However, this technique limits the data utility to the use of aggregate distribution. In another work, **(Groat *et al.* 2011)** used generalization and suppression technique to anonymize the data but not taken sensitive attributes into consideration disclosed the information. **Yuan *et al.* (2010)** suggested that the owners can define their privacy level by creating a taxonomy tree using generalisation. Owner's privacy is breached if an attacker is allowed to violate from sub-nodes and so this method is hard to implement.

A core problem in current approaches is that most popular PPDM algorithms concern about data stored in a centralized environment. Among most recent development in information and communication technologies, the distributed PPDM methodology have got to achieve immense amount of attention. Moreover, data hiding method have been dominated technique for privacy of an individual and do not consider the effect of data mining resulting in sensitive rules leakage. Due to these reasons, there is a need to extend the current tools and techniques into other problem domains or data mining tasks.

V.CONCLUSION

In this literature review, the taxonomy of PPDM, data mining, social networking privacy issues and machine learning have been explored in detail. Initially, discussed the Privacy preserving data mining (PPDM) along with their types. After that Social networking privacy was also covered. Additionally, the literature review section presented several frameworks and methods to develop the PPDM algorithm related to the privacy of social network users. In addition to all above, the problems in existing PPDM approach were addressed as well as issues in existing social networks privacy modelling. After the brief review of the literature, this research work is motivated on addressing the anonymization techniques using a classification model.

REFERENCES

1. Aghasian, Erfan, Saurabh Garg, and James Montgomery (2018). A privacy-enhanced friending approach for users on multiple online social networks. Computers 7.3: 42.

2. Amin Tootoonchian, Stefan Saroiu, Yashar Ganjali, and Alec Wolman (2009). Lockr: Better Privacy for Social Networks. In Proceeding of the 5th ACM International Conference on Emerging Networking Experiments and Technologies (CoNEXT).
3. Ben Niu, Xiaoyan Zhu, Qinghua Li, Jie Chen, and Hui Li (2015). A novel attack to spatial cloaking schemes in location-based services.In Future Gener. Comput. Syst. 49 ,125–132.
4. B. Zhou, J. Pei and W. Luk (2008). A brief survey on anonymization techniques for privacy preserving publishing of social network data. ACM SIGKDD Explorations Newsletter, vol. 10, no. 2, pp. 12_22.
5. Boyd, d (2004). Friendster and publically articulated social networking. In the Extended Abstracts of the Conference on Human Factors and Computing Systems (CHI 2004). Vienna, Austria, pp1279-1282.
6. Dolvara Gunatilaka (2011). A Survey of Privacy and Security Issues in Social Networks. In CSE571S: Network Security.
7. D. Rosenblum (2007). What anyone can know: The privacy risks of social networking sites. IEEE Security and Privacy, 5(3):40–49.
8. Dutton, W. H. (2004). Social transformation in an information society: rethinking access to you and the world. Retrieved from UNESCO Archives English - UNESCO HQ Social SciencesSSDCN(stock2E)-UNESCO.
9. Elena Zheleva, and LiseGetoor (2007). Preserving the privacy of sensitive relationships in graph data. In 1st ACM SIGKDD International Workshop on Privacy, Security and Trust in KDD (PinKDD 2007).
10. Erkin Z., Piva A., Katzenbeisser S., Lagendijk R., Shokrollahi J., Neven G., and Barni M. (2007). Protection and Retrieval of Encrypted Multimedia Content: When Cryptography meets Signal Processing. EURASIP Journal of Information Security, vol. 7, no. 17, pp. 1 - 20.
11. Fogel, J., andNehmad, E. (2009). Internet Social Network Communities: Risk Taking, Trust and Privacy Concerns. Computers in Human Behavior , 153-160.
12. Ganesan, M., and P. Mayilvahanan (2017). Cybercrime Analysis in social media Using Data Mining technique s." International Journal of Pure and Applied Mathematics 116.22: 413-424.
13. Gilbert Wondracek, Thorsten Holz, EnginKirda, and Christopher Kruegel (2010). Practical Attack to De-anonymize Social Network Users. IEEE Symposium on Security and Privacy, pp.223-238.
14. Gkoulalas-Divanis A, Verykios VS (2009). Exact knowledge hiding through database extension. IEEETransKnowlDataEng21(5):699–713.
15. Hai Jiang, Feng Shen, Su Chen, Kuan-Ching Li, Young-SikJeong (2015). A secure and scalable storage system for aggregate data in IoT. Future Gener. Comput. Syst. 49, 133–141.
16. Jian Wang, Yongcheng Luo, Yan Zhao, and Jiajin Le (2009), A Survey on Privacy Preserving Data Mining. In IEEE, First International Workshop on Database Technology and Applications.
17. J. Lin, Y. Cheng (2009). Privacy preserving itemset mining through noisy items.Expert Systems with Applications, vol. 36, pp. 5711-5717.
18. J. Kang (1998). Information privacy in cyberspace transactions. Stanford Law Review, 50(4):1193– 1294.
19. Kumar, G., and Kumar, K. (2012). The use of artificial-intelligence-based ensembles for intrusion detection: a review. Applied Computational Intelligence and Soft Computing, 21.
20. Kun Liu, Kamalika Das, Tyrone Grandison, and Hillol Kargupta (2008). Privacy preserving data analysis on graphs and social networks. In Next Generation of Data Mining, chapter 21, pages 419--437.
21. L. Backstrom, C. Dwork, and J. Kleinberg (2007). Wherefore art thou r3579x? anonymized social networks, hidden patterns, and structural steganography. In Proceedings of the 16th international conference on World Wide Web, pages 181–190. ACM.
22. L. Palen, and P. Dourish (2003). Unpacking" privacy" for a networked world. In CHI '03: Proceedings of the SIGCHI conference on Human factors in computing systems, pages 129–136, New York, NY, USA. ACM.
23. Martin Beck and Michael Marh¨ofer (2012). Privacy-Preserving Data Mining Demonstrator. In Proceedings of 16th International Conference on Intelligence in Next Generation Networks, IEEE.
24. M. M. Groat, W. Hey, and S. Forrest (2011). KIPDA: k-indistinguishable privacy-preserving data aggregation in wireless sensor networks. In Proceeding IEEE INFOCOM, pp. 2024–2032.
25. M. Yuan, L. Chen, and P. S. Yu (2010). Personalized privacy protection in social networks. In Proc. VLDB Endowment, vol. 4, no. 2, pp. 141–150.
26. Michael Hay, Gerome Miklau, David Jensen, Philipp Weis, and Siddharth Srivastava (2007). Anonymizing social networks. Technical Report 07-19, University of Massachusetts Amherst, Computer Science Department.
27. Murat Kantarcioglu, and Chris Clifton (2004). Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data.
28. Ning Xi, Cong Sun, Jianfeng Ma, Xiaofeng Chen, and Yulong Shen (2015). Secure service composition with information flow control in service clouds. Future Gener. Comput. Syst. 49, 142–148.
29. Nergiz ME, Gök MZ, Özkanlı U (2013). Preservation of utility through hybrid k-anonymization. In: Trust, privacy and security in digital business. Springer, Berlin, Heidelberg, pp 97–111.
30. Provost F, Dalessandro B, Hook R, Zhang X, and Murray A (2009). Audience selection for on-line brand advertising: privacy-friendly social network targeting. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '09, Paris, France.
31. Raymond Heatherly, Murat Kantarcioglu, and Bhavani Thuraisingham (2013). Preventing Private Information Inference Attacks on Social Networks. In IEEE Transactions On Knowledge And Data Engineering, Vol. 25, No. 8, pp 1849-1862.
32. Roy Ford, Traian Marius Truta, and Alina Campan (2009). P-Sensitive K-Anonymity for Social Networks. In Proceeding of International Conference on Data Mining,USA.

33. R. Gross, and A. Acquisti (2005). Information revelation and privacy in online social networks. In WPES '05: Proceedings of the 2005 ACM workshop on Privacy in the electronic society, pages 71–80, New York, NY, USA. ACM.
34. Shelke, Suchitra, and Babita Bhagat, Prof. (2015). Techniques for Privacy Preservation in Data Mining. International Journal of Engineering Research and. V4. 10.17577/IJERTV4IS100473.
35. Shaohua Wan, Hua Yang (2013). Comparison among Methods of Ensemble Learning. International Symposium on Biometrics and Security Technologies.
36. S. Lohiya, and L. Ragha (2012). Privacy Preserving in Data Mining Using Hybrid Approach. In Proceedings of Fourth International Conference on Computational Intelligence and Communication Networks, IEEE.
37. Sanaz Kavianpour, Zuraini Ismail, and Amirhossein Mohtaseb (2011). Preserving Identity of Users in Social Network Sites by Integrating Anonymization and Diversification Algorithms. In International Journal of Digital Information and Wireless Communications (IJDIWC), Hongkong, Vol. 1, Issue 1, pp 32-40.
38. S. R. M. Oliveira, and O. R. Zaıane (2010). Privacy preserving clustering by data transformation. Journal of Inf. Data Manage., vol. 1, no. 1, p. 37.
39. S. Weiss (2008). The need for a paradigm shifts in addressing privacy risks in social networking applications. In The Future of Identity in the Information Society, volume 262, pages 161– 171. IFIP International Federation for Information Processing.
40. Timm, Dianne T., and Duven, Carolyn J. (2008). Privacy and Social Networking Sites.
41. V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, Y. Theodoridis (2004). State-of-the-art in Privacy Preserving Data Mining. In SIGMOD Record, 33(1): 50-57.
42. Xueyun Li, Zheng Yan, and Peng Zhang (2014). A Review on Privacy-Preserving Data Mining.IEEE International Conference on Computer and Information Technology (CIT), 769 – 774.
43. X. Tang, and C.C. Yang (2010). Generalizing Terrorist Social Networks with K-Nearest Neighbor and Edge Betweenness for Social Network Integration and Privacy Preservation. In Proceeding of IEEE International Conference on Intelligence and Security Informatics.
44. Xuan Ding, Lan Zhang, Zhiguo Wan, and Ming Gu (2010). A Brief Survey on De-anonymization Attacks in Online Social Networks. In Proceeding of International Conference on Computational Aspects of Social Networks, Taiyuan, pp 611 - 615.
45. Yuan Cheng, Ravi Sandhu (2013). Preserving User Privacy from Third-party Applications in Online Social Networks. In Proceeding of 22nd international conference on World Wide Web companion, Geneva, Switzerland, pp 723-728.
46. Yan Zhu, Zexing Hu, Huaixi Wang, Hongxin Hu, Gail-Joon Ahn (2010). A Collaborative Framework: for Privacy Protection in Online Social Networks. In Proceeding of 6th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom), Chicago, IL, pp 1 – 10.
47. Zhang X, Liu C, Yang C, Dou W, Chen J (2013). Combining top-down and bottom-up: scalable sub-tree anonymization over big data using MapReduce on cloud.
48. Z. Lijie, and Z. Weining (2009). Edge Anonymity in Social Network Graphs. In Proceeding of International Conference on Computational Science and Engineering CSE, pp 108.