An Analytical Review of Privacy Preservation using K Anonymity Along with Bayesian Classifier in Data Mining

V Sravan Kumar
Research Scholar
Computer Science and Engineering,
Madhav University, Abu Road , Sirohi India,
sravan512@gmail.com

Dr. Rashmi Agarwal
Computer Science and Engineering,
Madhav University, Abu Road,
rashmiagarwalcse@gmail.com

Abstract— Privacy Preservation for social media is one of most trending research subject around the world. In addition to its trendiness, it is a very sensitive issue also. People around the world share their private information on the social media without thinking that it may affect their privacy. In such a condition it becomes the unrest duty to prevent the user information which is private. A lot of research workers have already put their ideas on table for the same issue. This paper studies the effect of Bayesian network in contrast to the prevention of the private data over social media. This paper also describes the pros and cons of using Bayesian Network for privacy preservation and also it compares some of the ethical prevention algorithms for the same. The evaluation has been done on the basis of ethical data mining parameters like Precision, Recall, F-Measure

***Keywords**— Privacy Preservation, Social Network, Bayesian Network*

*****

I. Introduction

Social media has shown tremendous growth in this era of technology. A number of parameters have contributed in the rapid growth in social media. The foremost involvement is of technology such as increased broadband availability, development in software tools and more powerful computers, emergence of more accessible mobile devices. The term "Social media" related to a wide range of Internet services which permit user to contribute in online interactions, exchange of diversified data, contribute in user-created communities, post their opinions on blogs and sharing of new stories. Social media allows different communities to form quickly and communicate efficiently. It encourages the involvement and feedback from everyone and blurs the line between the media and spectators. They are hardly any hurdles to accessing and making use of the content. Due to its ease of access, use and high speed people are engaging more in the online discourse where topics range from environment and politics to technology and the showbiz industry. Due to huge propagation of information on the network through large communities, the security of data is also an issue. Social media is unspoken without describing the Web 2.0: term defines a way in which user use the World Wide Web, a place where the discourse and content are continuously transformed by all the operators in cooperative way. Kalpan and Haenlein, describes social media as "A group of Internet based applications that build on the ideological and technological foundation of web 2.0 ". The basics form of social media which are services to users include the following:

1. Blogs: These are online journals in which the entries are in chronological order. They can be hosted free on websites as WordPress, Tumbir etc.
2. Wikis: It is a website that permits user to add or edit the information. It is website where participants can add or modify any page using their web browser. One known example is Wikipedia, a free online encyclopedia which uses wiki technology.
3. Podcasts: These are the audio and video files that are available by subscription like many paid tunes.
4. Social bookmarking: These sites permit users to organize and share links to websites. Example, reddit.
5. Social networking sites: Theses are web based service which allow individual to make a profile and connect with other participants within the bound network or system. Most popular are Facebook and LinkedIn.
6. Status-update services: These are mainly microblogging services, such as Twitter which allow people to share short updates about people, events or to check on other updates.
7. Media-sharing sites: It permit users to post videos and photographs such as YouTube, Instragram.

8. Virtual world content: These allow services like creating a virtual world in which user can interact.

A. *Privacy concerns of social media*

In the present time, hackers stalk the networks for victims. Hackers trick the users to use the shortened URLs created with bit.ly and visit into the damaging sites. Only One visit give chance to hackers to inject the virus in to the user's computer and mobile phones. Spyware are the software which hackers install into the users mobile, computer or iPad and they give the information about the user each activity on social media to the hacker. The only way to tackle with this issue is to never click on unwanted source or the source user is not aware of. Most of the network sites have information about the user's personal profile which hacker takes and tend to use them for breaking different account sites. Today, the main technique used by the hackers is by clicking on forgot password and then trying to recover password by email. Once they have access to email they can get users all information. Social users need to know once you post your personal information online, it is no longer private.

1. Social profiling and 3rd party disclosure: In this it is stated that no agency have right to disclose any kind of system record of any person to the 3rd party. Nowadays, Facebook asks for permission when any third party wants to access users in formation.
2. Companies: Some companies do not have clear privacy guidelines as twitter has scanned and imported the user's phone contacts on their websites to have information about their users.

So for protecting your password users need to crate strong password by including symbols, number and capital letters in it. Restrict your access of your personal information to the limited known ones and take full benefit of the different privacy options offered by social media sites like blocking the strangers. By installing antivirus, anti-spyware user can protect their information from virus and Trojans.

B. Privacy *preservation in Data mining*

Data mining is latest emerging field which includes databases, artificial inteligence and statistics. A large volume of data is collected by companies by which a useful information is extracted. on the basis of the on the demand need. Data mining is also known as "knowledge discovery". The main problem that arise in the mass collection of the data is its confidentiality. Privacy is the main key element for the companies statistical data and product information Privacy preservatiion is the main branch when data minining is concerned as it includes various methods as data-hiding, masking, suppression, aggregation, pertubation, anonymization. Privacy preservation in data mining is the prior conditionn for swapping confidential information like data analysis, validation and publishing. Rapid increase in internet phishing positioned a swere threat for the senstive information which is propogated over the web. It has created a wide security concerns on the users and enterprises worldwide. A lot of communication service provided through internet services: online banking, electroic commerce that exploit both human and software vulneraabilities suffer a lot of loss it can be financial or personal information. Therfore, privacy preservation data minining method are needed for secured and reliable information exchange over the internet.

These methods utilization reveal their ability to prevent the unfair use of data mining.

1. *Data distortion dependent PPDM*

It is method in which firstly the sensitive attributes are identified. The Threshold limit is set for the identification of the attributes. Using swapping techniques the data owner modifies the value under identified sensitive attributes. Modification of data is done in such a manner that the main data values are not changed.

2. *Association rule based PPDM*

In this techniques two probability parameters (fp and nfp) are employed for privacy preservation. Better accuracy is achieved by this method by tuning two parameters. When the fraction of frequent items among all the available items is less at that time output results are produced.

3. *Hide association rule based PPM*

In this technique, the correlation between the sensitive association rules (SAR) and each transaction in the database are examined by choosing the suitable item for modification. Association rule hiding is the technique of privacy preserving data mining to protect the sensitive association rules created by association rule mining.

C. K-anonymity

It is a privacy preservation technique for social networking data. While exchange of data, statistics or any company information on social media, they can have connection between them which make the data vulnerable to the 3rd party. When a specific person data is released with scientific guarantees so that the identification of the person remain anonymous. If the data possess k-anonymity property it means the degree of a particular group in the network be (k-1) elements. It means the information on the network is not vulnerable to loss. For example during the release of records they will change the main key and will replace them with dummy identifiers or we can add extra noise nodes to not make them identifiable to the intruders in the network. Even after adding dummy identifier, some set of attributes lead to identify breaches and are known as quasi –identifier. A set of attributes let birth date, zip code and gender attributes in the table can disclose the information of an individual and if attached so specific publically available information like voting list table lets to a leak of information on the network. K –anonymity prevents these type of privacy breach by confirming the release of record of individual if there is (k-1) distinct individuals whose related records are not separated from the previous.

The protection provided by k-anonymity is easy and simple to understand like if any table has k-anonymity for some value m than the one who knows the quasi identifier of someone cannot recognize the record equivalent to the individual. It provides against identity disclosure, it doesn't provide enough protection adjacent to attribute disclosure.

Because of the limitations of k-anonymity, l-diversity is introduced as the stronger notion of privacy.

D. Bayesian network

It is basically a graphical model for defining probabilistic relationship among set of variables. The conditional Independence relationship between the variables are encoded by Bayesian network. It provides a compressed representation of the joint probability distribution over the variables. In this domain is modeled by a list of variables ($X_{1,}$ $X_{2\ldots}$ $X_{n.).}$ Context about the problem domain is represented by a joint probability P($X_{1,}$ $X_{2\ldots}$ $X_{n.).}$. In this focused links defines casual direct influences. Every node has conditional probability table which quantifies the effects from the parents. There are no directed cycle's means it consist of directed acyclic graph.

Bayesian probability defines the degree of belief in particular event whereas classical probability deals with actual probability of an event.

II. RELATED WORK

In this paper,B.K. Tripathy and Anirban Mitra[1] has given the algorithm that follows the properties of k-anonymity and l-diversity during anonymisation. Basically, social media has three types of anonymisations. The first type has anonymisation of nodes. Anonymisation of the network structure is used in the second type. To get the best approch, third type i.e. the combination of first and second cases that mostly protects the node identification and the node's sensitive attribute values of nodes. The algorithms are implemented through small size social networks by means of graphs. The algorithms, authors have represented works on the higher degree of distance for the structural similarity of the nodes. To achieve anonymity, step by step computation of the partition of the clusters is taken place. Diversity for making the algorithm run efficiently is used, basically name as Algorthm. This algorithm can be extended by a multi-sensitive diversity. Alina Campan and Traian Marius Truta[2] have presented a generalization method for edges and calculate to count structural information failure. A greedy privacy algorithm is developed for anonymizing The social network, SaNGreeA (Social Network Greedy Anonymization) algorithm, is used that perform a greedy clustering processing to create a k-anonymous masked social network, specified an

Initial social network prototyped as a graph. Nodes from are describe by quasi-identifier and responsive attributes and edges from are undirected and unlabeled. Ninghui et al [3] have shown that the l-diversity with some limitations and proposed a new privacy notation called t-closeness. This approach has two parts, first that has the population in the released data and second for the specific individuals. The authors have used the Earth Mover Distance measure for t-closeness necessity. They have compared the efficiency and data quality of five privacy measures: k-anonymity; entropy -diversity; recursive- diversity; k-anonymity with t-closeness (t = 0.2); and k-anonymity with t-closeness 7 (t = 0.15). The comparison of the data quality of the five privacy measures by the discernibility metric and Minimal Average Group Size took place. Benjamin C. M. Fung et al [4] proposed a technique to k-anonymize a social network dataset with the objective of preserving common sharing patterns, which is the significant kind of knowledge necessity for marketing and consumer behavior analysis. They have conducted the experiment on three real-life datasets that are, and. For calculating the data efficacy on numerous patterns. They have measured the transformation of the common patterns before and after anonymization. A+ tool called MAFIA to take out the frequent patterns is also used. Bin Zhou and Jian Pei [5] have identified a vital kind of privacy attacks: neighborhood attacks. The authors have firstly identified the effect of the parameters used that are In the anonymization quantity measure. is settled first in the base. The number of values of are changed accordingly for measuring the number of edges added. The outcome shows that a person can trade off among adding edges and derive labels by alteration of the three parameters. Sri M.Vamsi Krishna et al [6] have put into practice together distinct l-diversity and recursive assortment. For attaining the requirement of k-degree l-diversity, the authors have designed a noise node with the algorithm to form a new graph from the original graph with the restraint of initializing the number of distortions to the original graph. The extensive experimental results illustrates that the noise node adding algorithms can grasp an improved result than the preceding work using edge editing only. Two privacy preserving data mining approaches emerged in recent years. The first protects data privacy using an extended role based access control approach where sensitive objects identification protects individual privacy. The second uses cryptographic techniques. A new solution integrating advantages of both techniques to reduce information loss and privacy loss was proposed by Vasudevan, et al [7]By using cryptographic techniques to store sensitive data and providing access to stored data based on individual's role, it ensured that data safety from privacy breaches. Privacy preserving distributed data mining has many applications each posing different constraints: What is meant by privacy, what the desired results are, how data s distributed, the constraints on collaboration and cooperative computing, were studied. Clifton et al [8] suggested that a solution to this was a components toolkit that can be combined for specific privacy-preserving data minimum applications. They also presented components of such a toolkit showing how they could be used to solve many

privacypreserving data mining problems. New tools for building PPDM techniques were developed and demonstrated new applications for this technology. There are still many challenges in this area, such as defining privacy constraints. As an example of the potential difficulties, imagine a scenario where the data mining results violate privacy. Secure multiparty computation definitions do not solve the problem. A specific PPDM problem was described by Zhan &amp; Du (2003): Company C wants to collect data from customers to form a data set for data mining. For data collection, C sends out a survey with a set of questions; each customer has to answer these questions and send back answers. But, as the survey has sensitive questions, not every user wants to disclose his/her answers to the questions; how could a method be developed so that C cannot learn a customer's actual answers, while being able to derive reasonably accurate data mining results? The randomized response technique was proposed for data collection. This method adds some randomness to answers to prevent data collectors from learning true information. To enhance privacy levels, a multi-group scheme was proposed where customers partition all answers into multiple groups, and for various groups, randomize data separately. Bertino et al [9] reviewed and summarized existing criteria and metrics in evaluating privacy preserving techniques. The aim of PPDM algorithms is to extract relevant knowledge from large data while protecting sensitive information. An aspect in design of such algorithms is identification of evaluation criteria and related benchmarks development. Recent research devoted effort to determine a trade-off between right to privacy and need of knowledge discovery. Often no privacy preserving algorithm exists that can outperform all others on all criteria. Hence, it is crucial to ensure a comprehensive view on metrics related to current privacy preserving algorithms to gain insights on how to design better measurement and PPDM algorithms. Basic paradigms and notions of secure multiparty computation and their relevance to privacy- preserving data mining was surveyed by Lindell&amp;Pinkas [10] . In addition to reviewing secure multiparty computation definitions and constructions, it discussed efficiency and demonstrated difficulties in constructing highly efficient protocols. It presented common errors prevalent in literature when secure multiparty computation techniques were applied to privacy-preserving data mining. Finally, it discussed relationship between secure multiparty computation and privacy-preserving data mining showing which problems it solved and which it did not. PPDM issues emerged globally. The recent PPDM techniques proliferation is evident. Motivated by increasingly successful techniques; new generation PPDM moves toward standardization. Oliveira et al [11] laid out what needs to be done and takes steps to propose such standardization: First, describe problems in defining which information is private in data mining, and discuss how data mining violates privacy. Then, based on users&#39; personal information and information concerning their collective activity, the privacy preservation in data mining is defined. Second, analyze implications of Organization for Economic Cooperation and Development (OECD) data privacy principles in a data mining context and suggest PPDM policies based on such principles. Finally, propose requirements to guide development/deployment of technical solutions. A brief overview of state-of- the-art in PPDM and current suggestions to proceed to PPDM standardization was summarized by Meints and Möllera [12] followed by how PPDM could be improved based on European Directive 95/46/EC, taking into account procedural/process related considerations. To illustrate them, financial sector scoring practice is used as example. Though this does not demonstrate all aspects relevant to data mining, it was analyzed from a recent data protection developments perspective. Additionally, with process chains containing basic data providers, service providers to calculate scoring values andbanks using mining results; the study analyzed requirements that data controllers had to meet. The issue of security violations when malicious parties provide false data was studied by Han &amp;Ng [13] The author identified secure scalar product protocols, 4 privacy vulnerabilities in many PPDM algorithms, proposing a general model of 2-party interaction. Its applicability to securely compute (x1+y1)(x2 + y2) and (x + y) log2(x + y) where xi and yi are private values held by each party respectively was demonstrated and it showed how the model could securely compute 4 commonly used kernel functions and other functions. The author also proposed 2 necessary conditions and 2 basic measures for adoption in the current malicious model. Huang &amp; Du [14]described a method to quantify privacy and utility. It then formulated quantification as estimate problems, and used estimate theories to derive quantification. An evolutionary multi-objective optimization method found optimal disguise matrices for randomized response technique. Experiments showed that the new scheme performed better than existing RR schemes. Jena et al [15] located an optimum balance between privacy and utility when publishing any organization's dataset. K means algorithm was used to cluster the dataset followed by k- anonymization. Privacy preservation is requirement to be satisfied and utility is a measure to be optimized.

References

[1] B.K. Tripathy, L.- Janaki, Jain Neha, &quot;Security against Neighbourhood Attacks in Social Networks&quot;, Proceedings National Conference on recent trends in softcomputing, pp. 216-223, 2009.

[2] Traian Marius Truta, Michail Tsikerdekis, and Sherali Zeadally, "Privacy in Social Networks," book chapter in "Privacy In a Digital, Networked World," ISBN: 978-3-319-08469- 5, pp. ?? – ??, Springer, 2015.

[3] Li, Ninghui; Li, Tiancheng; Venkatasubramanian, S. (April 2007). &quot;t-Closeness:Privacy Beyond k-Anonymity and l-Diversity&quot;. IEEE 23rd International Conference on Data Engineering, 2007. ICDE 2007:106–115.

[4] Benjamin C. M. Fung, Yan&#39;an Jin, Jiaming Li:Preserving privacy and frequent sharing patterns for social network data publishing. ASONAM2013:479-485.

[5] Bin Zhou and Jian Pei. Preserving Privacy in Social Networks Against Neighborhood Attacks. In Proceedings of the 2008 IEEE 24th International Conference on Data Engineering, ICDE &#39;08, pages 506-515, Washington, DC, USA,2008.IEEEComputerSociety.

[6] 1 Sri M.Vamsi Krishna, 2 Dr.K.V.V.S.Narayana Murthy, 3Ch.Srinu, "Method To Prevent Re-Identification Of Individual Nodes By Combining K-Degree Anonymity With L-Diversity", International Journal of Science Engineering and Advance Technology, IJSEAT, Vol 2,Issue11,November–2014.

[7] Rangarajan Athi Vasudevan , "A Novel Scheme for Secured Data Transfer Over Computer Networks", JUCS, 2011

[8] Clifton, C, Kantarcioglu, M, Vaidya, J, Lin, X, &amp; Zhu, MY 2002, 'Tools for privacy preserving distributed data mining', ACM SIGKDD Explorations Newsletter, vol.4,no. 2, pp. 28-34.

[9] Bertino, E, Fovino, IN, &amp; Provenza, LP, 2005, 'A framework for evaluating privacy preserving data mining algorithms', Data Mining and Knowledge Discovery,vol. 11, no. 2, pp.121-154

[10] Lindell, Y, &amp; Pinkas, B, 2000, 'Privacy preserving data mining. In Advances in Cryptology—CRYPTO 2000' Springer Berlin Heidelberg, pp

[11] Oliveira, SR &amp; Zaïane, OR 2004, 'Toward standardization in privacypreserving data mining', In ACM SIGKDD 3rd Workshop on DataMiningStandards,vol.7..36-54.

[12] Meints, M., &amp; Möller, J 2008, 'Privacy Preserving Data Mining'.

[13] Han, S, &amp; Ng, WK 2008,' Preemptive measures against malicious party in privacy-preserving data mining'. In SIAM International Conference on Data Mining,pp. 375-386

[14] Huang, Z, &amp; Du, W 2008, 'Optimizing randomized response schemes for privacy-preserving data mining', In Data Engineering, IEEE 24th International Conference on, IEEE, pp. 705-714.

[15] Jena, L, Kamila, N K, &amp; Mishra, S 2013, 'Optimizing the Convergence of Data Utility and Privacy in Data Mining', International Journal.