Hybrid and Decentralized Privacy Preservation using D-Anonymity and T-Closeness in Social Network

Annapurna Kattimani, Vijaylakshmi M., Channappa B. Akki

Abstract: ***Although Social Network (SN) knowledge is significant assets for data examination, freeing the data to the general public could reason an invasion of privacy. Privacy insurance is taken a lot of seriously than various data mining duties. The privacy problems are dealt with by several algorithms and strategies in the literature. But, perpetually there exists a trade-off between privacy and data.Our objective in this work is to design and develop a privacy-preserving solution for the social network. We have used K-anonymity and T-closeness algorithm and data anonymization. Further, data anonymization is decentralized by giving control of anonymization to the data owner. The solution is implemented on a dummy social network for testing the effectiveness of the privacy preservation solution proposed by us.***

Keywords: SN (social-networking), K-anonymity, T-closeness, Quasi-identifier.

I. INTRODUCTION

Nowadays, SNs are turning into an essential piece of day by day communications of current life. Facebook guarantees over 1.4 billion month to month and 900 million daily active users. SNs have obtained the quality of being widely admired in today's world. People use various SNs like Twitter, Facebook, Instagram, LinkedIn, etc.

An SN can be pictured as a chart in which the hubs are the people in the social network (or actors), and the links speak the connections or stream between the people [1].Users of SN store the personal sensitive information such as Location, relationship, education and many more, Further, SNs allow users creating and joining groups which have different interests, culture, attitude across the globe.The full publication of these personal data in its original form may attract adversary to hack the private data and use it for illegal things and thereby violating the privacy of the network users. We have chosen such an issue of privacy preservation in SN as our field of study.

Revised Manuscript Received on December 05, 2019.

Annapurna Kattimani, Asst. Professor, Dept. of Information Science & Engineering, ISE KLEIT Hubballi, Karnataka, India E-mail:anukattimani@gmail.com

Vjayalakshmi M., Associate Professor, School of Computer Science School of Computer Science and Engineering, KLE Technological University, Vidyanagar, HUBLI, Karnataka, India. E-mail:viju11@kletech.ac.in

Dr. Channappa B. Akki, Professor, Dept. of Computer Science & Engineering, IIIT, Karnataka, India. E-mail: akki.channappa@iiitdwd.ac.in

Privacy Types in Social Networks

Three categories in which the protection must be kept up are as follows[1]:

- Node Privacy: Client information need to be protected.
- Attribute privacy: User traits like interest, sex, area, etc ought to be guaranteed.
- Connection privacy: Information about relationship between the users must be preserved.

Types of Data in SN

Several types of social network data may be accrued from various belongings on the internet (i.e., various SN web sites) and extracted from the everyday activities and interactions among users[1]. In this context,

1. Explicit details the data which store clearly expressed information which is contributed by SN users, the explicit data includes text message, photos, videos, etc. here SN users enthusiastic participant in the formulation n of data.

a) Service data is a information which will be stored during creation of the account, to create account user must have to feed basic information about themselves which includes user's name, date of birth and country, etc.

b) Divulge data is data where another user of SN post on her/him SN profile like comments, captions posted entries, posted comments, shared links, etc.

c) Commend data is data where users post on other's profiles such as comments on other profiles, caption, tag, common links, etc.

d) Coincidental data is data where other SN user post about the Victim user it might consists of comments, notes, posted photos, etc.

2. Implied data is the set of data that is not clearly given or not explicitly provided by SN users. Here data is not provided by SN users through SN track all the information about users like the people we communicate and what post or video we often watch and SN analysis the behavior of the user. The data provided by Surmise originates from the examination of the behaviors of the users or is borrowed from explicit information about one or more users.

For precedent, the characteristics of relationships between multiple users can be predicted by analyzing the different aspects of communication patterns between users, e.g., instant memo, distributed photographs, and the quantity of normal companions.

a. Observable data is data that is generated from the user's behaviors. A social network can gather or gather data about the user's behavior and custom by stalking the arrangement of the bustle of the users and as a result, interpret the SN user's behavior. generated user behavior data can disclose diverse info such as with whom the SN users routinely interact, what the user routinely does on the SN, and in what information the SN user is implicated. These all such info is being collected by social networks by examining the articles that the SN users read, the game which they play on social networks, etc.

b. Derived data is the data that is not related to the behavior or habit of SN users. It is about the information that can be generated from all other data. For example, SN collects the IP address which can be used to construe the substantial location of users. If a user visits a different place in the city then one can say that SN users might live there as well.

II. PROBLEM IDENTIFICTION

As applied to the chosen topic of privacy preservation in SN, our objective is to design and develop a privacy preserving solution for SN. Further, we evaluate the proposed solution by deploying the same for a dummy SN.

III. RELATED WORK

The social network is a web-based application that gives numerous users to the interface, impart, connect and share the information on the web. There are distinctive social network sites like Facebook, Twitter, and LinkedIn, etc. These are used for interfacing the individuals and collaborating with one another.

People create personal profile information for various informal community locales to share their thoughts,

photographs, recordings, messages, texting and furthermore are also used for finding old buddies or finding human beings who have comparative interests or issues crosswise over various zones. Nowadays, social networks are persistently developing in number and size, the owners gather anmiraculous bulk of information about online social network users. The gathered data are in substance and furthermore, contain delicate information of the users attracting cyber-crimes. So the published data should not breach the privacy of the users.

Misuse of private information of an individual might be misused by the network users is called a privacy breach.

For the most part, network users have a solid discernment that the system administrators keep their private information secured. Social network operators are confronting a significant challenge to maintain online social network user's protection while publishing the social network data.

A vast amount of personal information that is voluntarily provided is disclosed in social networks and prone to misuse. The disclosure of information on the web is voluntary, many users are unaware of the risks that can occur (e.g. who can use their data and for what purposes). Stalking and Child exploitation is one of the initial privacy breaches that happened in 2010 on MySpace.In this case, minors were harassed and prompted for "age prerequisites and other wellbeing measures". In 2009, another crime was committed by Peter Chapman.Here, Peter acquired a character of someone to attract and assault a 17-year-old young lady.

To conclude, the increase in social networks has increased the cybercrime rate. Hence to prevent cybercrime, a major challenge for researchers is to provide a solution to this social problem**.**

Our identity is uncommonly important.Having the option to demonstrate you are huge for most pieces of life from applying for a home credit to getting recognizable proof. Having your identity stolen can bargain regular exercises and undermine your reputation, both secretly and expertly. That is the reason it's critical to ensure your protection on the web. Privacy can be achieved in two ways: proactive and reactive. In reactive, if someone misuses the information, corrective measures shall be initiated. Here the privacy has already been violated. Nowadays many fake accounts have been created just to violate social rules and/or to spoil the name of a person by accessing his (her) information. Such accounts can be taken down by authorized admin of social media if the victim or other user reports such an event. But if a victim fails to identify the fake account then fake account holder goes on misusing victim. In a proactive method, preserving privacy is achieved by some algorithm. This method makes user account information more secure and preserves privacy.

The issue of privacy preservation was addressed in several ways like cryptographic method in a dispersed habitat [2][3], data publishing, query acknowledge [4] and task-independent technique [5]. Cryptographic privacy preservation on dispersed information results in estimation impenetrableand also presupposes zero-knowledge proof [3]. The data publishing approach is much related to Query answering techniques, where in place of disseminating the data. The database claims queries as long as the answers do not infringement privacy [4]. The authors used K-anonymity, *L*-diversity and T-closeness algorithm on a medical dataset to preserve the privacy of users [4]. As per the study, most of the authors have used the available database for the approach followed, whereas we have used the real-time data in our study.

A. *k-anonymity*

K-anonymity is anviable protection saving procedure that has been first advised by Samarati and Sweeny [6] to avert associating a person to a record in an information table through a semi identifier QID.

A desk is known k- anonymity if for one file within the desk that has some value QID, in any event k-1 other facts also have the value QID.In a k-anonymity table, each record is indistinguishable from as a minimum k-1 other information concerning QID. An characteristic is called Sensitive, if the character isn't intrigued to uncover or an antagonist ought to now not be capable of know the cost of that characteristic.

k-anonymity can be executed in two ways: generalization and suppression. Generalization incorporates of rehabilitation (or documentation) a value with a less explicit but connotation reliable value.

Suppression includes veiling thevalue partially. k-anonymity ensures against identity revelation, it does not equip plentiful protection against attribute disclosure. This has been perceived by a few authors, e.g., [7, 8, and 9]. Two assaults were distinguished in [7]: the homogeneity attack and the background knowledge attack. This infiltrate clout the case wherever all the worth's for a sensitive value inside a group of k records area unit indistinguishable. In such cases, even supposing the info has been k-anonymized, the sensitive worth for the set of k records is also specifically foretold. Background Knowledge attack takes place when the adversary has well-known wisdom about the individual and with added logical analysis, an individual's sensitive attributes can be predicted. To cope with those barriers of k-anonymity, Machanavajjhala et al. [7] presented *l*-diversity as a more grounded idea of privacy.

l-Diversity

To overcome the disadvantage of K-anonymity the l-diversity is introduced. l-diversity contains at least one well-expressed value for sensitive attributes. Here grouping of more sensitive value will be done. Here 'l' represents the number of sensitive value that needs to be grouped in this strategy, the term ``well-represented" can be translated from multiple points of view. In this paper, the creators have utilized entropy as the data theoretic idea and thusly came the possibility of Entropy *l*-diversity.

Entropy l-Diversity:

The entropy of an equivalence class E is defined to be

$$\text{Entropy (E)} = -\sum_{s\in S}^{n} p(\text{E}, \text{s}) \log (p(E, s))$$

where$S$ is the set consisting of sensitive attributes, and $p(\text{E}, \text{s})$ is the records with fraction in E with asensitive values. Set S is divided into two different sets $S_a$andS_b then Entropy (S) >min (Entropy (S_a), Entropy (S_b)). A table subsist of entropy *l*-diversity for each commensurate class *E*, Entropy(E) ≥ log (l)

So, to accomplish*l*-diversity, entropy needs to be at least log (l) for the whole table. One of the exclusions for this circumstance is the sensitive attribute being extremely normal. Thus, this technique is prohibitive. At the same time l-diversity looks auspicious over k-anonymity, however, it has detriments too. This procedure is muddled [9]. Furthermore, l-diversity doesn't guarantee the aversion of credit exposure to a satisfactory level. It is slanted to skewness and comparability attacks.

B. *T-closeness*

In spite of the fact that k-anonymity and *l*-diversity outline a variety characterize an assortment of key conditions for security conservation, they are not adequate .Both these techniques do have a few downsides. As per [10], ``*l*-diversity is neither crucial nor satisfactory to anticipate characteristic revelation ". Consequently, a new technique is proposed in [10], called ``T-closeness" An proportionality class is said to have T-closeness provided the circulation of the quality in the entire table is close to a limit *T*anddistance between the distribution of a sensitive characteristic in this class.T-closeness decrease interrelationship between quasi-identifier and sensitive attributes.It protects against homogeneity attacks and perceive the linguistic proximity of attribute, a limitation of *l*-diversity.T-closeness constrain that the diffusion of a sensitive attribute in any eq. class is near to the dissemination of a sensitive attribute in the entire table.

IV. RESULTS AND DISCUSSIONS

Our literature survey revealed to us that the anonymization of data or adding up data was done by database administrator which might become a strong point of failure. Hence anonymization of data has to be decentralized. This can be achieved by giving control of the anonymization of data to the data owner. As per our literature survey, many authors have performed attribute anonymization without the user's concern, but privacy concerns will be changing from one user to other so we have come up with a solution**,** where the data owner decides about anonymization and level of privacy.

The data owner will be authorized to categorize his/her data into three groups: zero-sensitive, semi-sensitive, and full-sensitive. Based on the group of the data as categorized by the owner (zero-sensitive, semi-sensitive, and fully-sensitive), the anonymization technique can be selected. For example, if the owner categorizes the data as zero-sensitive, then the data need not be anonymized. If the owner categorizes a particular data as semi-sensitive, then the data can be anonymized using the K-anonymity algorithm. For the data categorized as Fully-sensitive, such data can be anonymized using the T-closeness algorithm. In traditional K-anonymity algorithm 'K' signifies the maximum number of attributes that would be anonymized. But in the proposed solution, categorization of data is with data owner he or she can declare any number of data as semi-sensitive, hence when we are applying K-anonymity 'K' loses its significance and 'K' will be varying from one user to other user and hence 'K' can be replaced by 'D' implying dynamic. Hence, traditional K-anonymity is renamed as D-anonymity in our work. The beauty of D-anonymity lies in the fact that D varies from user to user and also the control is not in the hands of database admin but lies with the data owner. The data which were classified as full-sensitive by the user,

T-closeness technique can be applied. In our study, we have designed a new T-closeness algorithm which will be discussed below. This entire solution is implemented using the Netbeans platform.

Our solution is tested on this platform. Our focus is not to develop a sellable product in this research study. However, we focus to develop a testbed for validating our proposed solution for privacy preservation.

A. *T-closeness algorithm*

T-closeness algorithm will be applied when user select full-sensitive option to anonymize the data, in our project we have considered T-closeness as highest level to anonymize, here we are performing eight step to anonymize the data.

Step1: Choose 4 private prime key p_1, p_2, p_3, p_4
Step2: Choose randomly two public keys h_1, h_2
Step3: Take ASCII of each character of the message to be encrypted a_1, a_2, a_3, a_4, $a_5 \ldots a_n$
Step4: Multiply each ASCII (step3) with prime key (step1) in round robin fashion.

$$(a_1 * p_1)(a_2 * p_2)(a_3 * p_3)(a_4 * p_4)(a_5 * p_5)(a_6 * p_6) \ldots \ldots \ldots (a_n * p_n) where\ 1 \leq i \leq 4$$

Store each result in variable 'r' i.e.

$$(a_1 * p_1) = r_1(a_2 * p_2) = r_2(a_3 * p_3) = r_3(a_4 * p_4) \ldots . (a_n * p_n) = r_n$$

Step5: Insert both the headers $h_1 h_2$ from step 2 on the left side of the result obtained in step 4.

$$h_1, h_2, r_1, r_2, r_3, r_4, r_5, \ldots \ldots \ldots . r_n$$

Step6: Subtraction operation is performed as below

$$h_1(h_2 - r_2), (r_1 - r_2), (r_2 - r_3) \ldots . (r_n - r_n + 1)$$

Store each result in variable 'c' i.e.

$$(h_2 - r_1) = c_1(r_1 - r_2) = c_2(r_2 - r_3) = c_3(r_3 - r_4) \ldots . r_n - r_n + 1 = c_n$$

Step7: Insert number of digits for each character in sequence to the result obtained by step 6

$$d_1 h_1, (d_2 c_1), (d_3 c_2), (d_4 c_3) \ldots . (d_n c_n)$$

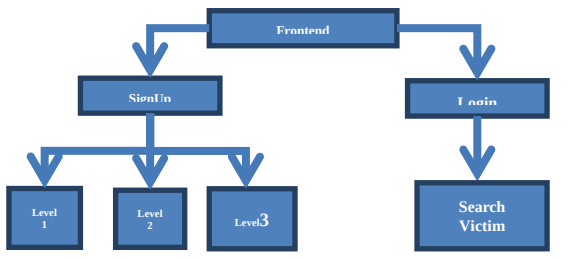
Step8: Insert sign bits according to sign ('1' for –ve and '0'for +ve).

Our above T-Closeness algorithm (designed by us) applied on few data categorized as full-sensitive by the owner and their anonymized data are shown Table 1 below:

Table 1: Data along with their equivalent data, anonymized as per our T-Closeness

Original data	anonymized data (Fully-sensitive)
Karnataka	3288032370354203415033880411910359702150 3558041140
Dharwad	3288031880366803117036770411050323403233
Hubballi	3288032160378302130339203987035090321603 381
India	328803223036990290034850411O6

V. IMPLEMANATION



The figure 1.1 shows the steps followed in social media privacy preservation

This project is implemented using the Netbeans platform. The frontend of our dummy social network named "friends-book" has an option to create an account (Sign up). The user needs to provide his/her basic information to proceed with the registration. During registration, the user needs to select the level of privacy for the attributes as shown in figure 1.1. These levels are used to anonymize the data. The data owner will be authorized to categorize his/her data into three levels: zero-sensitive, semi-sensitive, and full-sensitive. For example, if the owner categorizes the data as zero-sensitive, then the data does not need to be anonymized. If the owner categorizes a particular data as semi-sensitive, then the data can be anonymized using the K-anonymity algorithm. For the data categorized as Fully-sensitive, T-closeness algorithm can be used for privacy preservation. After categorizing the data the account will be created for the particular user. Once the account is created the user is ready to login to friends-book. After login when the victim is searched the anonymized data based on a level with which the account is created, few fields are masked rather than the original data so that the privacy of the user can be preserved.

VI. CONCLUSION

The privacy problems are directed by disticnt algorithms and methods in disquisition. But, perpetually there endure a trade-off between privacy and data. We have used K-anonymity and T-closeness algorithm for data anonymization. Further, data anonymization is decentralized by giving the control of anonymization to the data owner. We tend to conclude that our simple hybrid methodology undoubtedly competitive with the other complicated anonymization method in safeguarding the privacy. Additionally, there's no data misfortune. Furthermore any number of sensitive attribute can be taken care of.

REFERENCES

1. Madhurisiddula, yingshu li, "An Empirical Study on the Privacy Preservation of Online Social Networks" special section on privacy preservation for large-scale user data in social networks, 2018 IEEE Translations and content mining, VOLUME 6, 2018, DOI 10.1109/ACCESS.2018.2822693.
2. Benny Pinkas, "Cryptographic techniques for privacy preserving data mining", SIGKDD Explorations, Vol. 4, Issue.2, pp 12-19, 2002
3. Wenliang Du, Zhijun Zhang, "A Practical Approach to Solve Secure Multi-party Computation," in NSPW '02:workshop on New security paradigms, pp. 127-135, 2002.H. Poor, An Introduction to Signal Detection and Estimation. New York: Springer-Verlag, 1985, ch. 4.
4. Sheng Zhong "Privacy preserving algorithms for distributed mining of frequent item sets", *Intl. Journal oninformation sciences*, 177, 2007.
5. Poovammal, E., and M. Ponnavaikko. "Task independent privacy preserving data mining on medical dataset." *2009 International Conference on Advances in Computing, Control, and Telecommunication Technologies*.IEEE, 2009.
6. P. Samarati. Protecting respondent's privacy in microdata release. IEEE T. Knowl. Data En., 13(6): 1010-1027, 2001.
7. A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam.-diversity: Privacy beyond K-anonymity.In Proc. 22nd Intnl.Conf. Data Engg.(ICDE), page 24, 2006.C. J. Kaufman, Rocky Mountain Research Lab., Boulder, CO, private communication, May 1995.

8. T. M. Truta and B. Vinay. Privacy protection: p-sensitive K-anonymity property. In Proceedings of the 22nd International Conference on Data Engineering Workshops, the Second Intenational Workshop on Privacy Data Management (PDM'06), page 94, 2006.
9. X. Xiao and Y. Tao.Personalized privacy preservation.In Proceedings of ACM Conference on Management of Data (SIGMOD'06), pages 229–240, June 2006. (Basic Book/Monograph Online Sources) J. K. Author. (year, month, day). *Title* (edition) [Type of medium]. Volume(issue). Available: http://www.(URL)
10. N. Li, T. Li, and S. Venkatasubramanian, ``*t*-closeness: Privacy beyond *K*- anonymity and `l-diversity," in *Proc. IEEE Int. Conf. Data Eng. (ICDE)*, Apr. 2007, pp. 106_115.
11. Bourahla, Safia, and Yacine Challal. "Social Networks Privacy Preserving Data Publishing." *2017 13th International Conference on Computational Intelligence and Security (CIS)*. IEEE, 2017.
12. Siddula, Madhuri, Lijie Li, and Yingshu Li. "An empirical study on the privacy preservation of online social networks." *IEEE Access* 6 (2018): 19912-19922.
13. Hsu, Tsan-sheng, Churn-Jung Liau, and Da-Wei Wang. "A logical framework for privacy-preserving social network publication." *Journal of Applied Logic* 12.2 (2014): 151-174.
14. Talbi, Rania, Sara Bouchenak, and Lydia Y. Chen. "Towards Dynamic End-to-End Privacy Preserving Data Classification." *2018 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W)*. IEEE, 2018.
15. Brendel, William, et al. "Practical Privacy-Preserving Friend Recommendations on Social Networks." *Companion of the The Web Conference 2018 on The Web Conference 2018*. International World Wide Web Conferences Steering Committee, 2018.
16. Multi-kernel, deep neural network and hybrid models for privacy preserving machine learning.
17. Zhu, Hui, et al. "Efficient and privacy-preserving proximity detection schemes for social applications." *IEEE Internet of Things Journal* 5.4 (2018): 2947-2957.
18. Koranteng, Felix N., Isaac Wiafe, and Eric Kuada. "An Empirical Study of the Relationship Between Social Networking Sites and Students' Engagement in Higher Education." *Journal of Educational Computing Research*(2018): 0735633118787528.
19. Poovammal, E., and M. Ponnavaikko. "Task independent privacy preserving data mining on medical dataset." *2009 International Conference on Advances in Computing, Control, and Telecommunication Technologies*. IEEE, 2009.

AUTHORS PROFILE

Annapurna Kattimani (M.Tech)
Asst.Professor Dept. of Information Science &Engg
ISE KLEIT Hubballi -580030
Karnataka,India
Email:anukattimani@gmail.com

Vjayalakshmi M (M.Tech)
Associate. Professor School of Computer Science
School of Computer Science and Engineering
KLE Technological University, Vidyanagar,
HUBLI-580031Karnataka, India
Emai:viju11@kletech.ac.in
&Engg,KLETECH, Hubballi, Karnataka INDIA
Wireless Multimedia Networks Teaching Experience:20 years.

Dr. Channappa B Akki(Ph.D.)
Department of Computer Science and Engineering
Indian Institute of Information Technology, Dharwad-580029 Karnataka, India
Email: akki.channappa@iiitdwd.ac.in
Professor Dept. of Computer Science &Engg, IIIT,
Ph.D.– Electronics & Computer Engg, IIT (Roorkee)
AIMA Certified Manager (www.aima.in)
Life Member, IMAPS, LM-399 (www.imapsindia.org)