International Journal of Engineering & Technology, 7 (2.27) (2018) 40-45

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Enhancement of k-anonymity algorithm for privacy preservation in social media

Aanchal Sharma [1] *, Sudhir Pathak [1]

[1] *Chandigarh University, India*
**Corresponding author E-mail: aanchal08107@gmail.com*

Abstract

In recent times, more and more social data is transmitted in different ways. Protecting the privacy of social network data has turn out to be an essential issue. Hypothetically, it is assumed that the attacker utilizes the similar information used by the genuine user. With the knowledge obtained from the users of social networks, attackers can easily attack the privacy of several victims. Thus, assuming the attacks or noise node with the similar environment information does not resemble the personalized privacy necessities, meanwhile, it loses the possibility to attain better utility by taking benefit of differences of users' privacy necessities. The traditional research on privacy-protected data publishing can only deal with relational data and even cannot applied to the data of social networking. In this research work, K-anonymity is used for providing the security of the sensitive information from the attacker in the social network. K-anonymity provides security from attacker by making the graph and developing nodes degree. The clusters are made by grouping the similar degree into one group and the process is repeated until the noisy node is identified. For measuring the efficiency the parameters named as Average Path Length (APL) and information loss are measured. A reduction of 0.43% of the information loss is obtained.

Keywords: *APL; Cluster; Genetic Algorithm (GA); Information Loss; K-Anonymity; Social Media.*

1. Introduction

In recent years, with the great expansion of the Internet, social networks have become more and more popular. Now, social networking has taken a step toward mobility, which makes people more excited to share what they are doing in present time [1]. Mobile social networks also allow users to make group chats, play 'social games' or can communicate with users [2]. Due to these conveniences, more and more people use mobile social networks. According to the report of com Score, in the U.S., Instagram users spend more than 98% of their time using mobile devices instead of desktops, compared with 86%of users spend their time on twitter site [3].

Users interact with social media through web-based software/ web applications through computers, tablets, or smart phones and are typically used for messaging [4]. Social media is a device which is used by people to interact with friends and family in the starting time, but was later adopted by companies that wanted to use popular new communication methods to reach customers. The power of social media is to connect and share information with anyone or with multiple people on earth as long as people used social media [5]. Figure 1 shows an example of social media.

Fig. 1: An Example of Social Media.

With the enhancing computational power with large amount of data being integrated by different agencies are valuable for the individuals and has more risk in terms of privacy [6]. The datasets with information of individual has changed from existing data models to composite ones. The work in data prevention has similar trend and gives valuable solution for different data models [7]. The creative research field for protection of privacy in data publishing can be related relational data. The privacy protection in relational database has provided required research outcomes like K-anonymity, t-closeness and l-diversity etc. [8]. Though, protection of privacy for publishing data in social network is known as a novel issue, except, the privacy protection techniques for relational database could offer few ideas and few anonymization techniques and methods of privacy preservation of relational database that may enhance privacy protection in social network [9]. Various traditional anonymization techniques and models in social network are shown in Table 1. Simple anonymity is considered as simplest technology of anonymity, in which the node information for identification is restored by insignificant symbols like 1, 2...3

etc. Figure 2 below defines simple anonymity publishing in social networking [10]. This research has utilized k-anonymity privacy protection model that helps to protect the data. The description for the same is described in next section [11].

Table 1: Anonymization Methods in Social Network for Privacy Preservation

Model	Privacy	Against attacks	Anonymization policy
K-degree	Identification of edges and nodes	Node's degree attack	Graph construction/ greedy algorithm
K^2- degree		Friendship attack	Heuristic algorithm
d-neighborhood (d, k) anonymity		d-neighborhood	Clustering/Generalization
k-automorphism		some structural attack	Block copy/clock partition
K-isomorphism			Orbit copy/ automorphic division
k-symmetry			

2. K-anonymity in social networking

The model of k-anonymity was introduced by Samarati P and Sweeney L in 1998 for avoiding privacy revealing, that demands presence of some quantity of not recognized individuals in the data of publicized table that develops aggressor disable for distinguishing the existing privacy of an individual and secures individual privacy [12] [13].

Definition 1: K-degree anonymity can be described with the graph of social network that is G(V, E). In the defined graph, when every node has k- 1 edges. The degree of K-anonymity may oppose the opponent with the background node degree knowledge [10]. As shown in Figure 2, the data collection is defined by degree d = {2, 2, 2, 2, 1} in unique graph of social networking (a) and the degree collection is d={2,2,2,2,2} in two-anonymity social network graph (b) with the addition of one edge among 2 and 5 node. The described technique may develop all nodes with similar degree simply and will protect the privacy of social networking [14] [15].

Definition 2: K-neighborhood anonymity can be defined with graph, G= (V, E) in which there are k-1 nodes in which the neighborhood sub-graph consists of similar structure with node sub graph u for some nodes u. This technique may take opponent attack with background knowledge of neighborhood relationship [16].

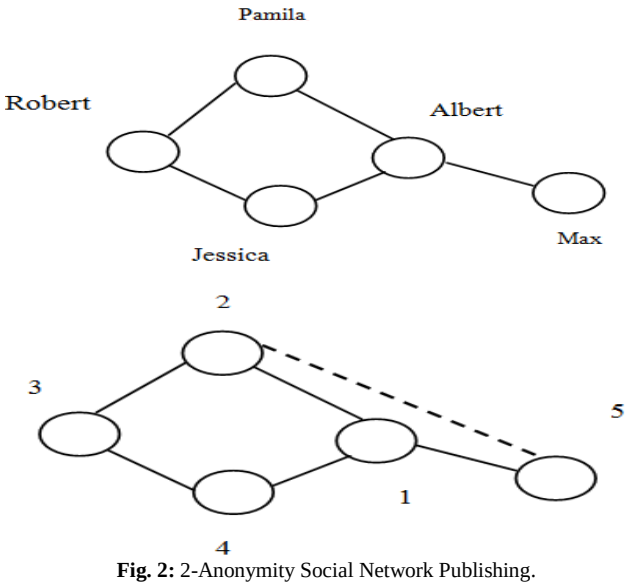

Fig. 2: 2-Anonymity Social Network Publishing.

3. Genetic algorithm

Genetic algorithm is an optimization approach used for the selection of appropriate solution. The genetic algorithm sustains population of n number of chromosomes along with proper fitness function [17] [18]. In the proposed work GA mainly used three functions named as mutation, crossover and selection function. In this research work GA is used for optimizing the noisy nodes that is added in the social network. This is done by optimizing the features of every node by using fitness function of Genetic algorithm [19].

Selection function: It is used to select parents, which take part at the next generation.

Crossover: In this process two parents are combines to produce a child for the upcoming generation.

Mutation: In this process changes are applied to every parent to form child. The algorithm of GA is defined below:

Algorithm 1: Genetic algorithm

```
Function GA ()
{
Initialize population;
Calculate fitness function;
While (fitness value != termination criteria)
{
Selection;
Crossover;
Mutation;
Calculate fitness function;
} }
end
```

4. Related work

Campan, A et al. [20] presented "Greedy algorithm" for the anonymization of social network. The authors also determined the loss of information in the procedure of anonymization during edge creation. SaNGreeA algorithm and the anonymization algorithm has been used which is dependent on collapsing clusters discovered from some classical k-anonymization algorithm for micro data. For the comparison among SaNGreeA algorithm and Zheleva's algorithm, the result quality being produced and computed by normalized generalization information loss and the normalized structural information loss. The experiments are calculated on 300 nodes for social network from Adult dataset form UC Irvine Machine Learning Repository. Six quasi- identifier attributes viz., work-class, age, race, native country, sex and marital status. The proposed algorithm is user balanced for preservation of more structural data of nodes or network values being attributed. This work lacks in extending the anonymity model for achieving the protection for social network attribute disclosure.

Zhou et al. [21] presented technique to preserve the social data privacy in case of attack occur in the network. If the enemy knows about any information between the neighbors, the attacker might be re-identified in the social network by using the process of traditional anonymization approach named as K- anonymity and l-diversity. The optimal computation of K anonymity and l diversity becomes a hard NP problem. The researcher has executed the anonymization method with real and synthetic datasets. The results are evaluated in four steps. In the first step, the researchers have shown the neighborhood attacks within real datasets that resulted to more serious problems of privacy. Subsequently, the examination of anonymization performance has been done with the amount of dummy edges, running time and label certainty penalty. Next, the researchers have evaluated the query performance for collective social queries with the anonymized data of social network. At last, the authors have inspected the diversity problems in social network data. Dataset of KDD Cup 2003 has been used for the examination. The research did not improve the anonymized social network data.

Zheleva et al. [22] focused the issues of link re-identification in the sensitive graph. The difficulty of assuming sensitive relations from the anonymized graph has been resolved. The author has shown the number of relationships being exposed at varied probable thresholds. For the data has been generated first that creates the data as per the data model. The input to the data generator has maximum number of nodes that are considered in the relationship, maximum amount of relationship that every student may have with each student. Later, the privacy preservation for anonymized data has been examined. The generation of data is from varied parameters, such as varied amount of classes and the amount of research groups among 10 and 30. Precision, recall rate with the amount of inferred sensitive relationship between the anonymized groups has been measured.

The precision of the technique has removed 50 % of the edge at arbitrary overlap with intact edge technique within sparse graph with almost dense graph overlap. The precision for two inhibited cluster edge techniques at k at 2 and at k at 6 also overlaps. The research has not considered the subtleties for the techniques effectiveness.

Bhaladhare et al. [23] proposed 'Greedy k-member' along with 'Systematic clustering' approach to reduce the loss of information. From the experiment it is determined that lesser information loss occur due to the use of systematic clustering technique. The attributes are utilized by the greedy and systematic approach during the generation of anonymized database. The disclosure risk has been resolved by using greedy technique whereas privacy preservation has been provided by using systematic clustering. Two approaches like Unequal combination of QI (quasi-identifier) and SA (sensitive attribute); Approach#2: Equal combination of QI and SA are considered for the examination. Each Method is considered for the applications of medical database as the private sensitive information like some disease could reveal when the mining of medical database takes place. Adult database has been used from UCI machine learning repository for the observation. The considered dataset has 32561 records with 15 attributes. The experiment has been analyzed on varied k-values like 20, 40, 60, 80 and 100. Parameters like total information loss and the execution time are measured for the execution. It has been concluded that Approach: 1 that is Unequal combination of QI and SA; Approach: 2 that is Equal combination of QI-SA has less information loss that traditional clustering approaches like Greedy k-member algorithm and Systematic clustering algorithm. This research can also considered the hybridization of different SA and QI attributes.

Kefei Mao et al. [24] proposed a verification system for privacy preservation in the field of healthcare system. The protection has been provided by using the smartcard. This method provides user safety from different attacks. Four operational stages have been considered for the experimentation, like system parameter generation, transmission and authentication with password update stage. The researchers have used IBE (Identity-based encryption) and IBS (Identity-based signature) for the schemes. A significant proposal has been provided by utilizing quadratic residue statement. The security can be enhanced by identity based cryptography method.

V. Kishor and S. B. Shrimantrao [25] applied slicing scheme in Hadoop. Slicing method is helpful to protect Medical & Government information. The presented plan demonstrates that the slicing provide better usability than 'simplification system and bucketization technique'. The researcher has designed one mapper class with one reducer class for executing Hadoop operation. The work has been executed by using Slicing method by using four nodes within HDFS. Anonymization method can be used in the future to anonymized the data. Tuple grouping algorithm can be used for dispersing the bucket accurately. Slicing method in cloud environment can be implemented and could be compared in Hadoop.

5. Problem formulation

A social network graph can be characterized as:

A social network is designed by using 4 tuple defined as :
G (a, b ,c, d) here, 'a' is the vertex set, b is the edge set, 'c' is the interconnection group and 'd' is the connection, which changes as per the dataset choice used in the simulation.
For the secure communication, a social network may be defined as a graph G (a, b, c, d).
Figure 3 below defines a graph of social network comprises of six number of vertex and edge.

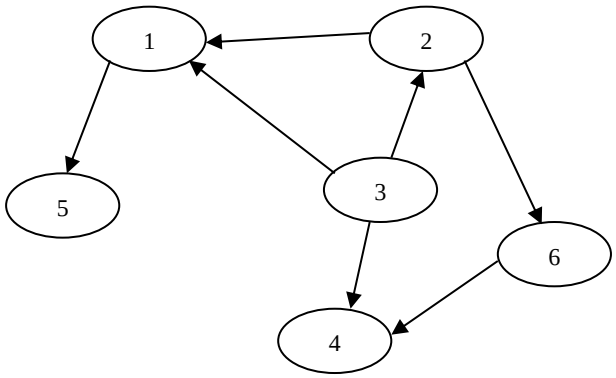

Fig. 3: Graph G (A, B).

The degree table for the above graph is shown Table 2. From the Table 2 it is clear that the vertex 1, 2 and 3 belongs to the similar degree, vertex 4 and 5 also belongs to the similar degree but vertex 6 has different degree. The vertex 1, 2, 3 belongs to group 'A' and vertex 4 and 5 belongs to group 'B' by using the concept of K-anonymity. Table 3 shows grouping according to K-anonymity.

Table 2: Degree of Every Node for G(A,B)

Node	In- degree	Out-degree	Total degree
1	2	1	3
2	1	2	3
3	0	3	3
4	1	0	1
5	1	0	1
6	1	1	2

Table 3: Grouping According to K-Anonymity

vertex	In- dgree	Out-degree	Total degree	
1	2	1	3	Group-A
2	1	2	3	
3	0	3	3	
4	1	0	1	Group-B
5	1	0	1	
6	1	1	2	

6. Methodology

The flow of research work is shown in Figure 4 below.
Step 1: Initially, a graph is made for the social network with vertex 'a' and edge 'b.
Step 2: Depends upon the connections made in the network, groups are created. The vertex with the same degree comes in one group and the vertex with different degree comes in the other group.
Step 3: The group A have degree with higher connection whereas group B has lower connection accuracy. Thus it becomes necessary to differentiate among higher and lower degree connections.
Step 4: If the link among the nodes are disconnected then there is an indication of attack coming in the network. Thus, to provide the security to the network a security table has to be maintained.
Step 5: The process of reconstructing the group is continued until the secure architecture has been obtained.

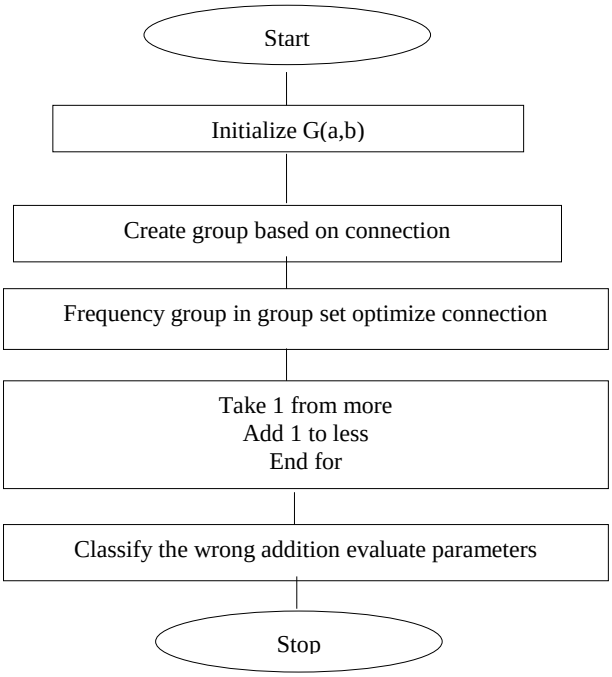

Fig. 4: Flowchart of Proposed Work.

Table 4: Input of the Proposed Work

Main Node	Node Connection Value	Connecting Node	Connection Weight	Weight Label
1	37171	0	1	1
1	37171	2	1	1
1	37171	8	1	1
1	37171	11	1	1
1	37171	19	1	1
1	37171	22	1	1
1	37171	23	1	1
1	37171	24	1	1
1	37171	25	1	1
1	37171	26	1	1
1	37171	27	1	1
1	37171	28	1	1
1	37171	29	1	1
1	37171	30	1	1
1	37171	31	1	1
1	37171	32	1	1
1	37171	33	1	1
1	37171	34	1	1
1	37171	35	1	1
1	37171	36	1	1
1	37171	37	1	1
1	37171	38	1	1
1	37171	39	1	1
1	37171	40	1	1

The input data used for the proposed work is shown in Table 4 above. The above table comprises of five columns having main node, node connection value, connection node, connection weight and weight label. The first step is to make K-anonymity graph, which is formed by grouping the nodes in such a manner that the nodes with same degree comes under same group. The degree of graph can be increased by creating edges among nodes. Noise nodes are added to know the accuracy of the designed framework. After obtaining the label for every noise edge, the decision has been taken as per the allocated edge labels and the edge connectivity in the graph. For an example, for a noise node "N" let us consider the label on the nearby edges to N are: {u1, u2, u3……..um}. The selection of the rule has been decided as per the formula written as shown in equation 1 below.

$$\max\left(\sum_{i=1}^{m}\sum_{j=1}^{u} N_{ocon}\left(B_{v,j}, B_{vei}, j, um\right)\right) \quad (1)$$

Here, vei represents the connecting node with v through edge I, B_{vei}, j is vei's label on j dimension.

N_{ocon}- is the edges in the actual graph in which Bv,j number of labels are connected withB_{vei}, j through an edge with ui labels.

7. Experimental results

To know the efficiency of the proposed work, the parameter named as average path length for small as well as for big data is measured. The simulation is performed in MATLAB simulator. The average path length (APL) is described below:
The average path is used to measure the connection among the two labels. Let two levels are defined as L1 and L2. Let us assume an un-weighted graph 'G' comprises of 'A' number of vertices. Let G(a1,a2), where a1,a2ϵA indicates the smallest distance between (b/w) a1 and a2. Let us assume that G(a1,a2)=0 if a2 does not reached from a1, then the APL can be written mathematically as shown in equation 2 below.

$$APL_G = \frac{1}{a\times(a-1)} \times \sum_{i\neq j} G(a_i, a_j) \quad (2)$$

Here, a signifies the number of vertices in graph 'G'

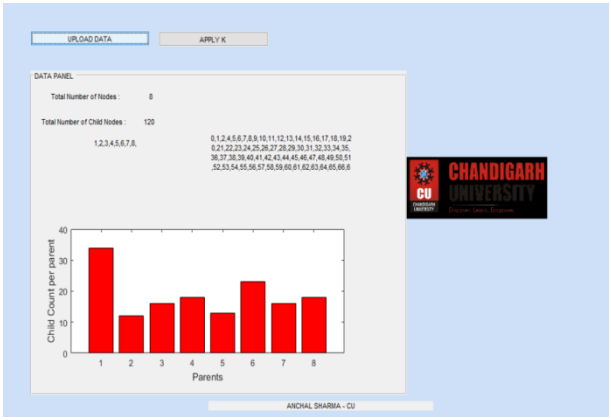

Fig. 5: Child Count per Parent.

Figure 5 defines the GUI along with the child count per parent for small data. The GUI of the proposed architecture includes two buttons named as upload data and apply- K. The data panel displays total number of 8 nodes along with 120 number of child nodes. Here, x-axis represents the 8 number of parent nodes and y-axis defines the child count for every 8 nodes individually.

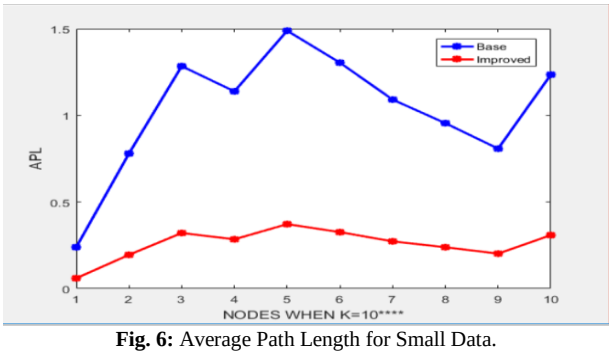

Fig. 6: Average Path Length for Small Data.

Figure 6 defines the APL of the proposed research work. Here x-axis signifies the nodes with K=10, y-axis signifies the APL values. The value of K is changes from 1 to 10. The comparison of proposed with the existing work has been shown. Red line represents the proposed work APL values whereas blue line represents the existing work graph. From the above graph it is concluded that the APL of proposed work is less as compared to the existing work which means that the graph creation and privacy is better than the existing work.Table 5 shows information loss for proposed and existing work.

Table 5: Information Loss for Proposed and Existing Work

K	Information loss of existing work [23]	Information loss of proposed work
1	33	32
2	33.3	33.1
3	33.5	33
4	33.6	33.2
5	34	33.8
6	34.2	33.1
7	35	34.6
8	35.2	34.1
9	36	35.8

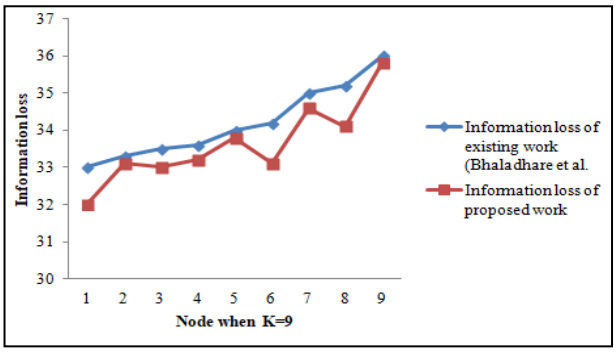

Fig. 7: Information Loss for K=9.

Figure 7 depicts the information loss for K=9. The comparison between proposed and existing work [16] are represents by red and blue line. From the above graph, we observe that the loss of information during grouping the clustering is less as compared to the existing work, which shows better clustering. The average information loss measured for the proposed and existing work is 33.63 and 34.2 respectively. There is an improvement of 0.43% has been observed for the proposed work.

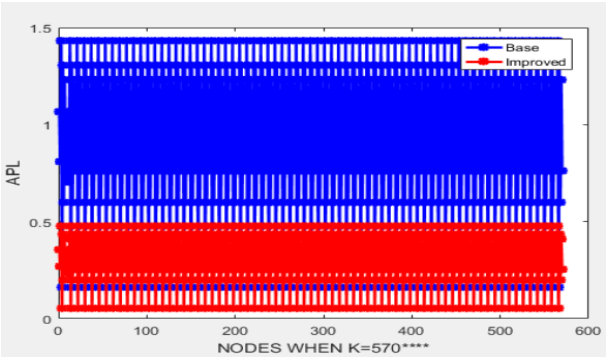

Fig. 8: Average Path Length for Big Data.

Figure 8 represents the APL measured for big data. Here the range of K lies from 0 to 600. For the existing work the APL is high and the maximum value of APL measured is 1.5. For the proposed work The APL measured for the big data is 0.5, which shows higher accuracy and privacy.

8. Conclusion

In this research, we have proposed an anonymization scheme for social media data. The concept of K-anonymity has been used for providing safety of the social media network. The noisy node has determined by using the concept of clustering. The privacy has been provided for small as well as large social data. The average values of APL measured for the existing and proposed work are1.06 and 0.26 for small data and for large data; the average values for existing and proposed work are 1.5 and 0.5 respectively. There is a reduction of 0.43% of the information loss from existing work. In future to enhance the privacy preservation of proposed work, classification techniques such as ANN (Artificial neural network), SVM (Support vector machine) can be used.

References

[1] Pham, V. V. H., Yu, S., Sood, K., & Cui, L. (2017). Privacy issues in social networks and analysis: a comprehensive survey. *IET Networks*.

[2] Gerbaudo, P. (2018). *Tweets and the streets: Social media and contemporary activism*. Pluto Press.

[3] Keküllüoğlu, D., Kökciyan, N., &Yolum, P. (2016, August). Strategies for privacy negotiation in online social networks. In *Proceedings of the 1st International Workshop on AI for Privacy and Security* (p. 2). ACM. https://doi.org/10.1145/2970030.2970035.

[4] Zhang, S., Li, X., Liu, H., Lin, Y., &Sangaiah, A. K. (2018). A Privacy-Preserving Friend Recommendation Scheme in Online Social Networks. *Sustainable Cities and Society*. https://doi.org/10.1016/j.scs.2017.12.031.

[5] Xu, L., Jiang, C., Chen, Y., Wang, J., &Ren, Y. (2016). A framework for categorizing and applying privacy-preservation techniques in big data mining. *Computer*, *49*(2), 54-62. https://doi.org/10.1109/MC.2016.43.

[6] Francis, J., & Stokes, M. (2012). *U.S. Patent No. 8,140,502*. Washington, DC: U.S. Patent and Trademark Office.

[7] Jung, E. K., Levien, R. A., Lord, R. W., Malamud, M. A., Mangione-Smith, W. H., &RinaldoJr, J. D. (2012). *U.S. Patent No. 8,203,609*. Washington, DC: U.S. Patent and Trademark Office.

[8] He, Z., Cai, Z., & Yu, J. (2018). Latent-data privacy preserving with customized data utility for social network data. *IEEE Transactions on Vehicular Technology*, *67*(1), 665-673. https://doi.org/10.1109/TVT.2017.2738018.

[9] Li, J., Yan, H., Liu, Z., Chen, X., Huang, X., & Wong, D. S. (2017). Location-sharing systems with enhanced privacy in mobile online social networks. *IEEE Systems Journal*, *11*(2), 439-448. https://doi.org/10.1109/JSYST.2015.2415835.

[10] Xiao, X., Chen, C., Sangaiah, A. K., Hu, G., Ye, R., & Jiang, Y. (2017). CenLocShare: a centralized privacy-preserving location-sharing system for mobile online social networks. *Future Generation Computer Systems*. https://doi.org/10.1016/j.future.2017.01.035.

[11] Wang, Y., Cai, Z., Chi, Z., Tong, X., & Li, L. (2018). A differentially k-anonymity-based location privacy-preserving for mobile crowdsourcing systems. *Procedia Computer Science*, *129*, 28-34. https://doi.org/10.1016/j.procs.2018.03.040.

[12] Diwakar, A. K., Singh, N. K., &Tomar, D. S. (2017, January). End user privacy preservation in social networks against neighborhood attack. In *Asia Security and Privacy (ISEASP), 2017 ISEA* (pp. 1-9). IEEE. https://doi.org/10.1109/ISEASP.2017.7976991.

[13] Wei, R., Tian, H., & Shen, H. (2018). Improving k-anonymity based privacy preservation for collaborative filtering. Computers & Electrical Engineering. https://doi.org/10.1016/j.compeleceng.2018.02.017.

[14] Fei, F., Li, S., Dai, H., Hu, C., Dou, W., & Ni, Q. (2017). A K-Anonymity Based Schema for Location Privacy Preservation. IEEE Transactions on Sustainable Computing. https://doi.org/10.1109/TSUSC.2017.2733018.

[15] Liu, P., Bai, Y., Wang, L., & Li, X. (2017). Partial k-Anonymity for Privacy-Preserving Social Network Data Publishing. International Journal of Software Engineering and Knowledge Engineering, 27(01), 71-90. https://doi.org/10.1142/S0218194017500048.

[16] Liu, X., Xie, Q., & Wang, L. (2017). Personalized extended (α, k) anonymity model for privacy preserving data publishing. Concurrency and Computation: Practice and Experience, 29(6). https://doi.org/10.1002/cpe.3886.

[17] Wang, H., Huang, H., Qin, Y., Wang, Y., & Wu, M. (2017). Efficient Location Privacy-Preserving k-Anonymity Method Based on the Credible Chain. ISPRS International Journal of Geo-Information, 6(6), 163. https://doi.org/10.3390/ijgi6060163.

[18] Zhang, S., Li, X., Liu, H., Lin, Y., & Sangaiah, A. K. (2018). A Privacy-Preserving Friend Recommendation Scheme in Online Social Networks. Sustainable Cities and Society. https://doi.org/10.1016/j.scs.2017.12.031.

[19] Amardeep singh, Divya bansal,& sanjeev sofat (2014) A Privacy Preserving Techniques in Social Network Data Volume 87 - No.15

[20] Campan, A., &Truta, T. M. (2009). Data and structural k-anonymity in social networks. In *Privacy, Security, and Trust in KDD* (pp. 33-54). Springer, Berlin, Heidelberg https://doi.org/10.1007/978-3-642-01718-6_4.

[21] Zhou, B., & Pei, J. (2011). The k-anonymity and l-diversity approaches for privacy preservation in social networks against neigh-

borhood attacks. *Knowledge and Information Systems*, *28*(1), 47-77. https://doi.org/10.1007/s10115-010-0311-2.

[22] Zheleva, E., &Getoor, L. (2008). Preserving the privacy of sensitive relationships in graph data. In *Privacy, security, and trust in KDD* (pp. 153-171). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-78478-4_9.

[23] Bhaladhare, P. R., & Jinwala, D. C. (2016). Novel Approaches for Privacy Preserving Data Mining in k-Anonymity Model. J. Inf. Sci. Eng., 32(1), 63-78.

[24] Kefei Mao, Jie Chen, Jianwei Liu and Mengmeng Wang, 'Security enhancement on an authentication scheme for privacy preservation in Ubiquitous Healthcare System,' 2015 4th International Conference on Computer Science and Network Technology (ICCSNT), Harbin, 2015, pp. 885-892.

[25] A. V. Kishor and S. B. Shrimantrao, 'Performance enhancement and analysis of privacy preservation using slicing approach over hadoop,' 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, 2016, pp. 353-357.