RESEARCH ARTICLE

SAMRIDDHI Volume 14, Issue 2, 2022 Print ISSN: 2229-7111 Online ISSN: 2454-5767

Graph-based Mechanism to Prevent Structural Attack over Social Media

Jitendra Patel, Ravi K. S. Pippal*

Department of Computer Science, RKDF University, Bhopal, Madhya Pradesh. India

Abstract

Social media sites contain the personal information of the users, which entice the attackers. The attacker performs different types of attack on the social media site to get the users sensitive information. User privacy may be breached as other passive and active attacks are performed on social media sites; to prevent such a scenario, the network operator releases the data anonymized. Social media operators fetch and store social media users' data to share among various third-party consumers. The network operator releases the complete graph in anonymized and sanitized versions because the fetched information often contains sensitive data. But it does not provide a full guarantee of user privacy. This paper proposed a solution that provides a neighborhood adjacency matrix-based anonymization process for the social network graph. This anonymization process may be used to counter the neighborhood attack over the social network graph. The proposed anonymization process increases the number of isomorphic neighborhood networks by adding dummy edges in the social network graph. Therefore, a user may not be re-identified in a social network graph based on their unique neighborhood.

Keywords: Adjacency matrix, Graph Anonymization, Graph-based Attack, Neighborhood Attack, Social Media, Social Media Mining.

SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology (2022); DOI: 10.18090/samriddhi.v14i02.00

Introduction

On the social network, end-users use the social media platform to share views, knowledge, information, and mutual interaction. End-users willingly provide their personal and private information for profile creation over social media sites, including phone number, profile picture, relationship status, email etc.

User interaction on social media generates rich data that contains sensitive information, e.g., users' attitudes towards any local and global political, clinical,[1,2] environmental, critical, and inflammable issues. Social media operators share user-generated content with third-party for research and data analytics, such as advertisement companies, political parties, and manufacturing companies for revenue collection. As user-generated content contain sensitive information, social media operator share anonymized graphical information.[3,4]

Suppose the advisory has background knowledge of the social media structure. In that case, the anonymized graphical information of social media may be vulnerable in the sense of structure-based attacks such as subgraph and neighborhood attack.[5,6] With structure-based attacks, an adversary can re-identify the targeted user node from the published social network.[7]

Recently, neighborhood anonymization of the social network restricts the adversary with background knowledge of neighborhood structure to prevent structure-based attacks.[8-10] This paper presents the anonymization of social network data and preserves user privacy against neighborhood attacks.[11,12] The neighborhood attack based on a user and its neighbour's information identifies the isomorphic structure.[13-15] If two or more neighborhood networks are isomorphic in the social network graph, and then the adversary cannot place a unique vertex neighborhood sub-network.[16,17] The proposed methodology increases the isomorphic neighborhood network in the social network graph by adding established imitation relationship edges.[18]

In this paper, a counter solution to prevent Neighborhood attacks has been presented. It has been found that if two or more neighborhood networks are isomorphic in social network graph, it is difficult for attackers to re-identify a specified vertex from the released anonymous graph. Also,

Corresponding Author: Ravi K. S. Pippal, Department of Computer Science, RKDF University, Bhopal, Madhya Pradesh. India, e-mail: ravesingh@gmail.com

How to cite this article: Patel, J., Pippal, R.K.S. (2022). Graph-based Mechanism to Prevent Structural Attack over Social Media. *SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology*, 14(2), 1-7.

Source of support: Nil

Conflict of interest: None

by simply adding dummy edges in the anonymous graph to avoid vertex re-identification will lead to greater information loss. So the proposed solution is based on increasing the isomorphic neighborhood network in the social network by adding minimum number of dummy edges. The detail of large social network data and implementation details of the 1-neighborhood adjacency matrix on the big data environment is present in the second and third sections.

Proposed Method

The user's data provided by social networking sites is anonymized, which means that a meaningless identifier replaces the user's critical information. Recently, anonymization procedures have been used to add dummy edges and corners to protect social networks' graphic data. However, the inclusion of false corners and edges increases the noise level in anonymous data.

This paper focuses on reducing the loss of information by lowering the noise level added for anonymization. Minimize the degree of noise by optimizing the number of false edges. An optimized artificial edge can maintain isomorphic groups and reduce information loss. Therefore, data from social networks remained helpful to researchers and data analysts.

This paper proposes a neighborhood adjacency matrix-based anonymization (NAMA) process for an extensive social media network, as shown in Figure 1. In NAMA, an optimized number of dummy edges are added to a unique sub graph; however, the number of vertices remains the same. Unique sub-graph and vertices are highly vulnerable to structure-based attack. NAMA approach has different modules for anonymized uniqueness of sub-graph and vertices.

For anomonization of uniqueness of vertices, it initially extracts a set of vertices degree in increasing order, as-

$$v = \{v_i^d, v_j^d, \ldots\ldots, v_m^d\} \tag{1}$$

Where u_V is the set of vertices with superscript their degree and v_i^d represent vertices v_i having degree d. However, if degree (d) of any v_i is unique for all v_i in v, then v_i is vulnerable vertices v_V.

$$d(v_i) = \begin{cases} unique \ \forall \ valnerable \ v_i \\ common \ \forall \ safe \ v_i \end{cases} \tag{2}$$

After that numa approach calculate identical degree vertices count (idvc), i.e., summarized all the vulnerable vertices (v_V) as:

$$idvc = \sum_{i=1}^{n} v_i \tag{3}$$

If idvc of v_V is greater than one, then the NAMA approach adds a dummy edge among v_V, else if idvc of v_V is one then add the dummy edge with vertices having more than two idvc until $d(v_V)$ does not become common.

However for anomonization of the uniqueness of sub-graph, NUMA approach extract vertices having identical degrees as vertices belong to vulnerable sub-graphv_{sg}. Next extract the 1-Neighborhood networks and create adjacency matrices. Now sort the adjacency Matrices. From the sorted vertices, vertices having a higher adjacency matrix have been selected. Subtract the lower adjacency matrix from higher adjacency matrix. The resultant adjacency matrix shows the required dummy edges for anonymization. Thus, dummy edges are added in that vertex neighborhood network that has lower adjacency matrix. After adding dummy edges the vertices neighborhood networks become isomorphic, as shown in Figure 2.

Algorithm1 [Stepwise explanation of proposed solution]

1. Start
2. Select the un-anonymized social network graph G (V, E) from the database, where V = $\{v_1, v_2, v_3, \ldots, V_n\}$ is set of

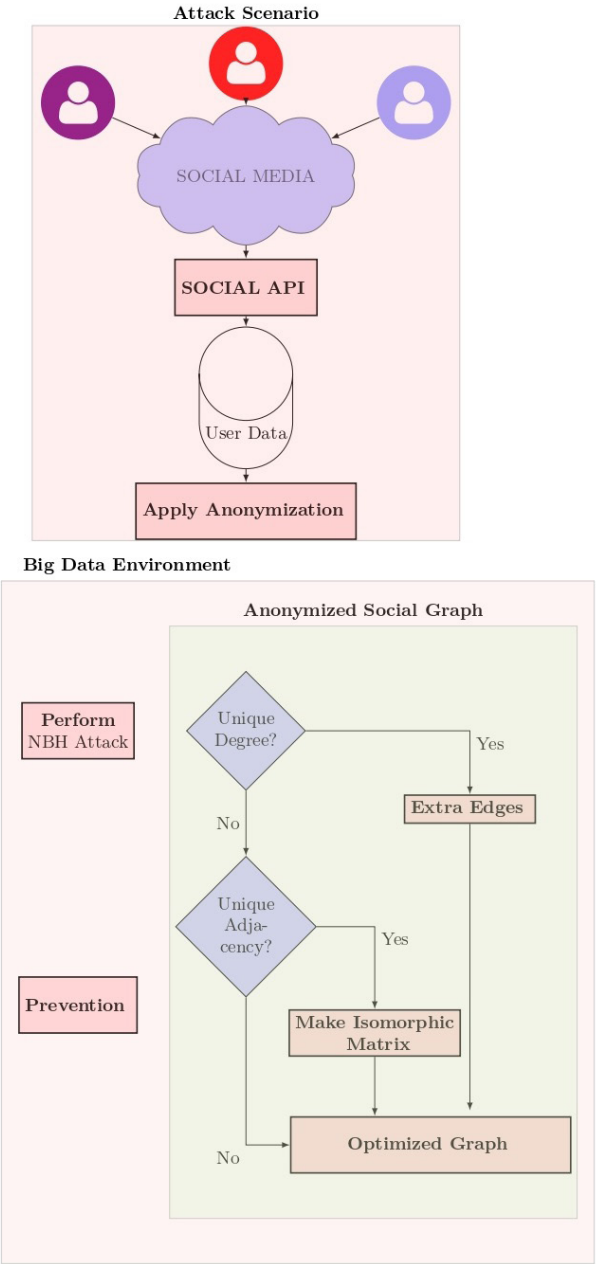
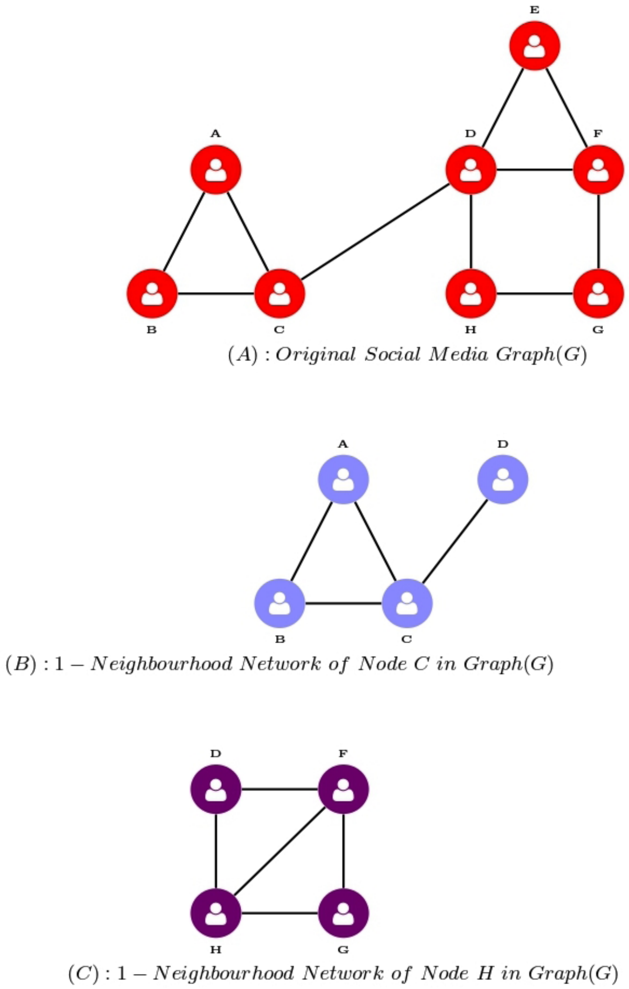


Figure 1: Proposed framework to identify neighborhood attack

n vertices and E = {e_{ij}= (v_i,v_j) |v_i,v_j in V , i not equal to j} is the set of edges between the n vertices of graph G. The degree of a node is decided by number of edges connected to with it. The set of degree D = {d1,d2, dn} is the available degrees in G.

3. Maintain a list of degree di in that all of the degrees from G, those appear once in G.
4. A vertex-degree list is created; in this list all of the vertices V and their degrees are present.
5. Extract the neighborhood networks N(vi) from G. Here, N(vi) is the set of neighborhood networks N(vi) = {v1,v2,…. vs} in the graph G.
6. The adjacency matrices Av for all N(vi) is created. It maintains a list of adjacency matrices.
7. Now select a degree from degree list di, created in step 2.
8. Select the vertices from the vertex-degree list those degree equal to the di, vertex-degree list created in step 3.
9. Select the adjacency matrices of all vertices those degree di.
10. Check the cost of all adjacency matrices and sort it according to the increasing order of cost.
11. Now calculate the difference between the all alternate adjacency matrix one after the other and store the result in AR, result adjacency matrices.
12. Extract the edges from AR, update the set of dummy edges ED by inserting edges information and go back to step 4 for next degree.
13. Remove the duplicate edges from the set of dummy edges ED and union the set of edges E of G and the set of dummy edges ED and generate a new anonymized set of edges E' = {E union ED}.
14. Update set of edges E with anonymized set of edges E' in the social network graph G.
15. Now the social network graph G becomes anonymized social network graph G'.
16. End.



(A) : Original Social Media Graph(G)

(B) : 1 – Neighbourhood Network of Node C in Graph(G)

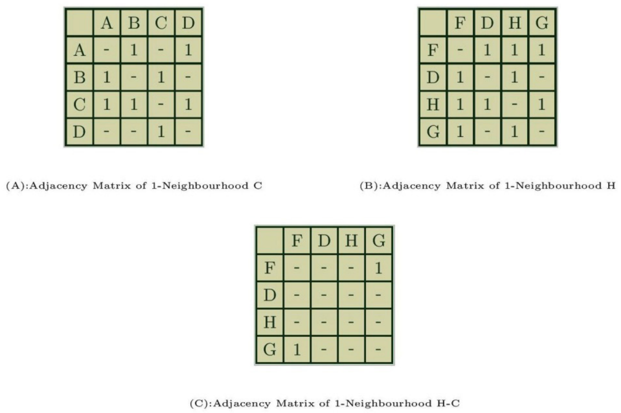
(C) : 1 – Neighbourhood Network of Node H in Graph(G)

Figure 2: Comparing of neighborhood networks before anonymization

The original social network is shown in Figure2(a). The 1-neighborhood networks of vertex C and vertex H is shown in Figure 2(b) and 2(c), respectively. Vertex C and vertex H has the same degree, but their neighborhood networks are different.

First, select a degree in social network graph G. Here, degree 3 is selected, the vertex C and vertex H has degree 3. Now, the 1-neighbor subgraph for both vertices is created. In first subgraph vertex A, vertex B, and vertex D are connected with node C and also node A, B are also connected with each other. Same as in second subgraph, vertex D, vertex F and vertex G are connected with node H. Vertex F , vertex G are connected with each other and also there is a connection between vertex D , vertex F. Both nodes C, H has the same number of neighbors but neighborhood networks are not isomorphic. The comparison can be done through adjacency matrices to find out the difference between neighborhood subgraphs, as shown in Figure 3.

To find the difference between networks with one neighborhood of node C, H, adjacent matrices are created and sorted in descending order. Here the node C adjacency matrix is higher than the node H adjacency matrix. Thus, the vertex C adjacency matrix is subtracted from the vertex



	A	B	C	D
A	-	1	-	1
B	1	-	1	-
C	1	1	-	1
D	-	-	1	-

(A):Adjacency Matrix of 1-Neighbourhood C

	F	D	H	G
F	-	1	1	1
D	1	-	1	-
H	1	1	-	1
G	1	-	1	-

(B):Adjacency Matrix of 1-Neighbourhood H

	F	D	H	G
F	-	-	-	1
D	-	-	-	-
H	-	-	-	-
G	1	-	-	-

(C):Adjacency Matrix of 1-Neighbourhood H-C

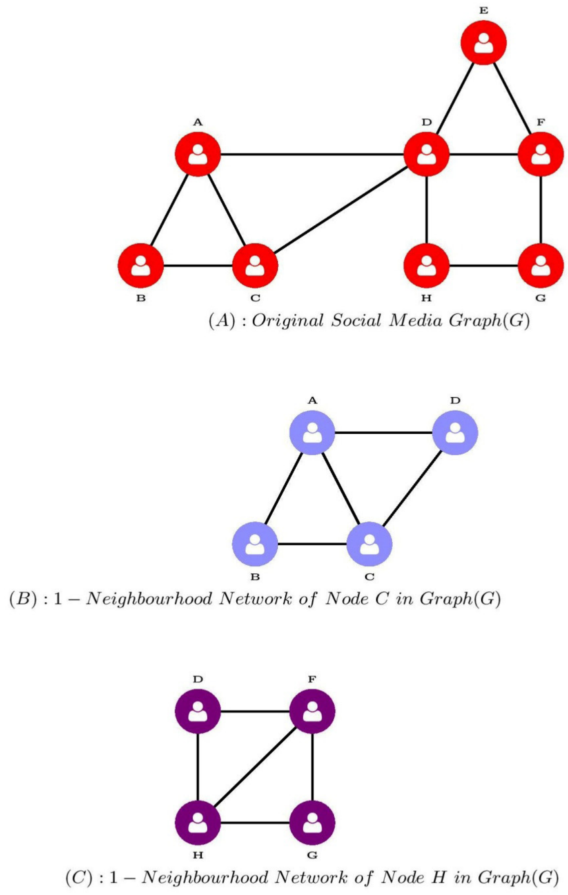
Figure 3: Adjacency matrix

H adjacency matrix. The resulting adjacent matrix C 'shows the required dummy edge between vertex D and vertex A to anonymize the neighborhood network of node M 1-neighborhood network. After adding dummy edges to the social network graphic G, it changes to the anonymized social network G. The 1-neighborhood network of M node and N node is isomorphic in the anonymized social network G '.

The anonymized social network graph G 'is shown in Figure 4(a). After adding dummy edges to the original social network chart, the similarity between the vertical C and the vertex H1 neighborhood networks is shown in Figures 4(b) and 4(c).

Experimental Setup

In the social big data network anonymization, all the experiments were conducted on a system running the Ubuntu 16.04 operating system, with a 2.3 GHz Core(TM) i3 CPU, 3.0 GB RAM and a 320 GB hard drive with open-source software Neo4j (version 3.3.5) and R (version 3.3.0). The program is implemented in R programming language. This paper illustrates the comparative analysis of social network graph anonymization on different sets of two different datasets i.e. Real-time Twitter dataset, and Gnutella Peer to Peer Network Dataset.



$(A):Original\ Social\ Media\ Graph(G)$

$(B):1-Neighbourhood\ Network\ of\ Node\ C\ in\ Graph(G)$

$(C):1-Neighbourhood\ Network\ of\ Node\ H\ in\ Graph(G)$

Figure 4: Comparison of neighborhood networks after anonymization

Real-time Twitter Dataset:

Three set of real-time data is extracted/crawled from Twitter. Firstly, 100 tweets is crawled, the network has 315 vertices and 435 edges. The data is collected from the followers of Modi. Each node represents the followers and edges represent a connection between a pair of hosts. Secondly, 1000 tweets have been selected. The network has 2313 nodes and 4110 edges. Thirdly, 10000 tweets is extracted, and the network has 20350 vertex and 43500 edges.

Gnutella Peer to Peer Network Dataset

The Gnutella peer to peer network dataset represents a directed graph where the edge represents a connection between a pair of Gnutella hosts. It has 22687 nodes and 54705 edges.

Result Evaluation

The anonymization process tampers the originality by pouring some noise over the network, i.e., adding or removing the edges in published data. However, the level of anonymization depended upon the degree of noise added to the networks. The analysis is carried out on common parameters, including the number of nodes, the number of actual edges, the number of dummy edges, and the average degree of the graph. In this paper, social network anonymization experiments were performed on real-time Twitter and Gnutella peer-to-peer network datasets, as shown in Table 1.

The graphical description of anonymized social network graph, i.e., a number of vertices, the minimal number of dummy edges required to be added, the total number of resultant edges and an average degree, are shown in Table 2. The change in network information before anonymization and anonymization is visible after comparing Tables 1 and 2. The number of vertices has remained unchanged. Whereas after incorporating the dummy edge, network density increases, reflecting both in the total number of edge count and average degree.

Evaluation of the NUMA anonymization approach is carried out by using the following evaluation metrics:

a. Anonymization Reflection of Average Degree(γ^{ad}): γ^{ad} is the reflection of the change in average degree before and after the anonymization through numa, as shown in equation 4.

Table 1: Parameters of unanonymized social network graph

Dataset	*#Tweets*	*# Vertices*	*#Edges*	*#Average Degree*
Twitter Dataset	100	315	435	3
	200	554	1085	4
	1000	2313	4110	4
Gnutella Dataset	501	501	710	3
	1001	1001	2068	4
	5001	5001	15798	6

Table 2: Parameters of anonymized social network graph

Dataset	*#Tweets*	*#Vertices*	*#Dummy Edges*	*#Total Edges*	*#Average Degree*
Twitter Dataset	100	315	101	536	4
	200	554	251	1336	6
	1000	2313	2513	6623	5
Gnutella Dataset	501	501	189	899	4
	1001	1001	958	3026	5
	5001	5001	4568	20366	8

Table 3: Evaluation of Anonymized Social Network Graph

	Tweet	*RRAD*	*RRAE*	*Noise*
Twitter Dataset	100	0.33	0.23	7.5
	200	0.50	0.23	4.4
	1000	0.25	0.61	1.3
Gnutella Dataset	501	0.33	0.27	4.9
	1001	0.25	0.46	2.3
	5001	0.33	0.29	0.2

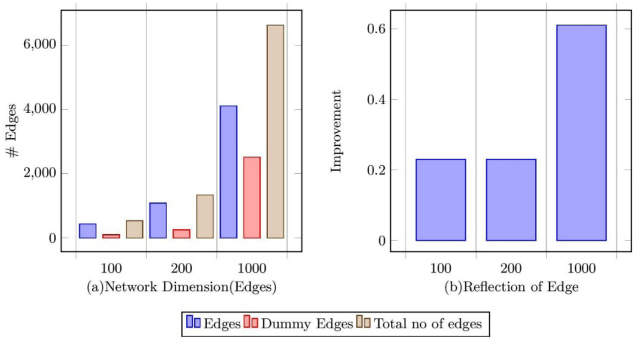


Figure 5: Reflection of edges after anonymization over twitter dataset

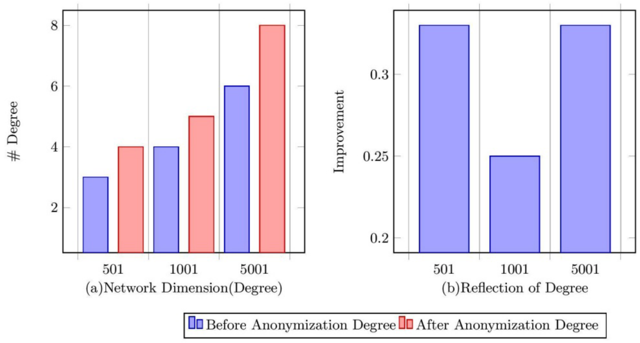


Figure 6: Reflection of degree after anonymization over twitter dataset

$$\gamma^{ad} = \frac{\sum_{i=1}^{n} d(v_i) \in G' - \sum_{i=1}^{n} d(v_i) \in G}{\sum_{i=1}^{n} d(v_i) \in G} \tag{4}$$

Here, G and G' represent the graph before and after the anomonization.

b. Anonymization Reflection of Edge (γe): $\gamma e$is the reflection of the changed total edge number after anonymization, as shown in equation 5.

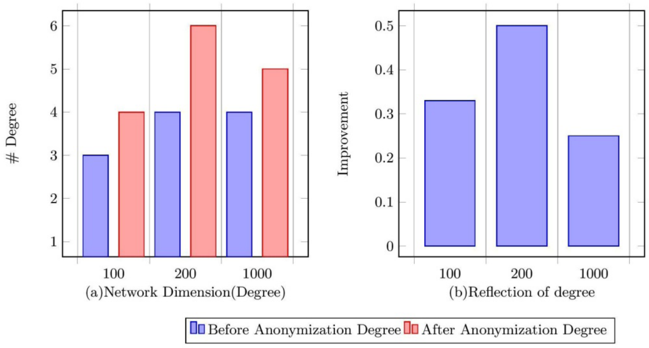


Figure 7: Reflection of degree after anonymization over gnutella dataset

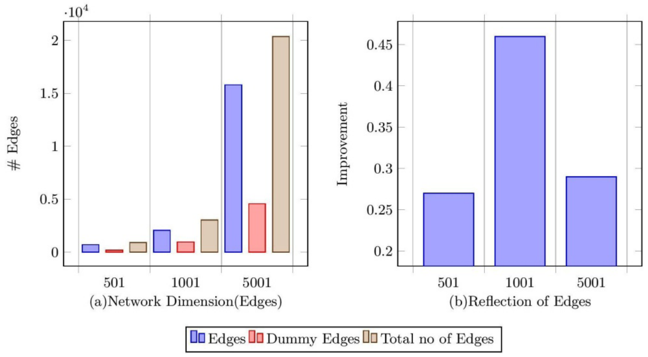


Figure 8: Reflection of edge after anonymization over gnutella dataset

$$\gamma^{e} = \frac{\sum_{i=1}^{n} e_i \in G' - \sum_{i=1}^{n} e_i \in G}{\sum_{i=1}^{n} e_i \in G} \tag{5}$$

Noise Level(α^l): Noise in the network may be created due to changes in network parameters i.e., number of vertices, edge, and average degree. α^l is measured as the total reflection in network information after anonymization, as shown in equation 6.

$$\alpha^{l} = \frac{\gamma^{v} + \gamma^{e} + \gamma^{ad}}{|v| + |e| + |ad|} \tag{6}$$

Average Degree of graph before and after anonymization is shown in Figure 5 and Relative ratio of average degree is shown in Figure 6. The relative ratio of average degree over Twitter dataset is measured between 1.65–1.75. and the relative ratio of average degree in Gnutella peer to peer

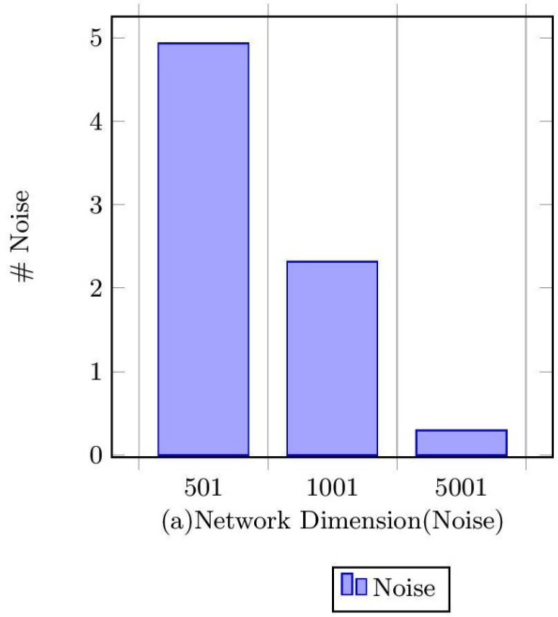


Figure 9: Reflection of noise after anonymization over Gnutella dataset

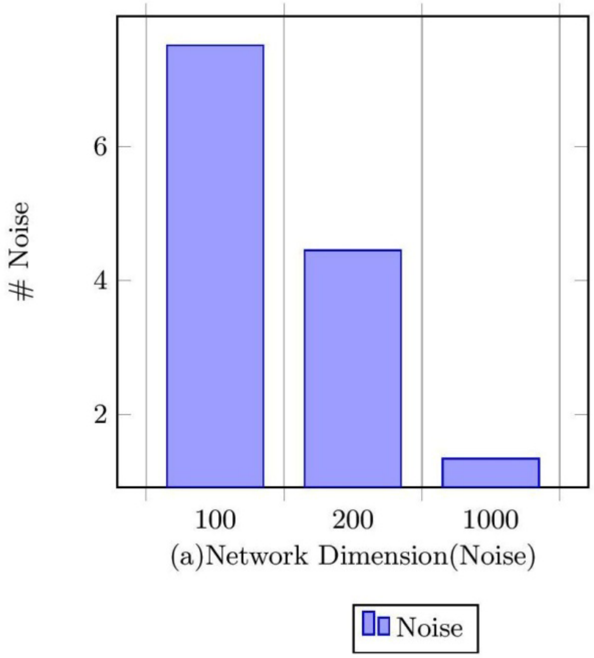


Figure 10: Reflection of noise after anonymization over Twitter dataset

network is measured between 1.3 - 1.75. The changes of a few edges or vertices using adjacency-based anonymization still have a small effect on the average degree.

Edge Change of graph before and after anonymization is shown in figure 7 and Relative ratio of edge change is shown in figure 8. The relative edge change ratio over the Twitter dataset is measured between 0.05 and 0.2. and the relative ratio of edge change in Gnutella peer to peer network is measured between 0.3 and 0.85. In big social networks, this adjacency matrix-based anonymization changes a small portion of vertices and edges without significantly affecting the neighborhood.

The noise level rise after anonymization in the social network's graph is shown in Figures 9 and 10. The noise level is measured between 0.05 to 0.75 and 0.08 to 0.3, respectively, in the anonymized Twitter and Gnutella Peer to Peer Network dataset.

The proposed approach decreases the noise level over social networks. Hence, it reduces information loss. Adding a dummy edge in the social network graph increases the social network's isomorphic nature. That leads to an increase in the privacy-preserving level of social network data. But simultaneously, these dummy edge pore some noise over social networks. Noise over social network data affects the results of those experiments conducted on social network data. The presented anonymization process acquires a marginal noise level, i.e., approximate 0.05\%-0.8\%, by adding and subtracting the optimized dummy edge.

Conclusion

Social media has become a significant part of human's day to day life. Social media provides a easy medium to connect with family and friends for communicating and sharing. People use social media environments to interact with old companions, maintain relationships, or even meet new friends, thus strengthening the overall connectivity among social media user. Social media sites contain the users' personal information, which entices the attackers. The attacker performs different types of attack on the social media site to get the user's sensitive information. User's privacy may be breached as a different type of passive and active attacks are performed on social media sites. To prevent such scenario network operator releases the data in an anonymized form. Recent anonymization process to preserve the use of social network graph data to add dummy edges and vertices. The addition of dummy corners and edges increases the noise level, which can cause information loss. The information loss is directly proportional to the number of dummy edges and the vertex added. Information loss is still a problem in social network anonymity. Including dummy corners and edges in social network data can change the social network graph's originality and increase the noise level. If excessive noise levels are presented in anonymous social network data, researchers and data analysts may receive an unfair result.This paper presents a neighborhood adjacency matrix-based anonymization process to counters the neighborhood attack over the social media data. NUMA increases the number of isomorphic neighborhood networks by adding dummy edges. Any user may not be re-identified in a social network graph based on its unique neighborhood network.

References

[1] Abawajy, J.H., Ninggal, M.I.H., Herawan, T. (2016). Privacy preserving social network data publication. IEEE Communications Surveys Tutorials 18(3), 1974–1997.

[2] Jamil, A., Asif, K., Ghulam, Z., Nazir, M.K., Mudassar Alam, S., Ashraf, R., Mpmpa, (2018). Amitigation and prevention model for social engineering based phishing attacks on facebook. In: 2018 IEEE International Conference on Big Data (Big Data), 5040–5048.

[3] Ji, S., Li, W., Gong, N.Z., Mittal, P., Beyah, R. (2016). Seed-based de-anonymiz ability quantification of social networks. IEEE Transactions on Information Forensics and Security 11(7), 1398–1411.

[4] Ji, S., Li, W., Srivatsa, M., Beyah, R. (2016). Structural data de-anonymization: Theory and practice. IEEE/ACM Transactions on Networking 24(6), 3523–3536.

[5] Ji, S., Mittal, P., Beyah, R. Graph data anonymization, de-anonymization attacks, and de-anonymize ability quantification: A survey. IEEE Communications Surveys Tutorials 19(2), 1305–1326 (2017).

[6] Kergl, D. (2015). Enhancing network security by software vulnerability detection using social media analysis extended abstract. In: 2015 IEEE International Conference on Data Mining Workshop (ICDMW), 1532–1533.

[7] Liu, G., Wang, C., Peng, K., Huang, H., Li, Y., Cheng, W., Socinf, (2019). Membership inference attacks on social media health data with machine learning. IEEE Transactions on Computational Social Systems 6(5), 907–921.

[8] Ninggal, M.I.H., Abawajy, J. (2011). Attack vector analysis and privacy-preserving social network data publishing. In: 2011 IEEE 10th International Conference on Trust, Security and Privacy in Computing and Communications, 847–852.

[9] Orabi, M., Mouheb, D., AlAghbari, Z., Kamel, I. (2020). Detection of botsin social media: A systematic review. Information Processing and Management 57(4), 102250.

[10] Patil, N.A., Manekar, A.S. (2015). A novel approach to prevent personal data on a social network using graph theory. In: 2015 International Conference on Computing Communication Controland Automation, 186–189.

[11] Rekha, H.S., Prakash, C., Kavitha, G. (2014). Understanding trust and privacy of big data in social networks - A brief review. In: 2014 3rd International Conference on Eco-friendly Computing and Communication Systems, 138–143.

[12] Reza, K.J., Islam, M.Z., Estivill-Castro, V. (2017). Social media users' privacy against malicious data miners. In: 2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), 1–8.

[13] Sharma, V.D., Yadav, S.K., Yadav, S.K., Singh, K.N., Sharma, S. (2021). An effective approach to protect social media account from spam mail – A machine learning approach. Materials Today: Proceedings.

[14] Sushama, C., Sunil Kumar, M., Neelima, P. (2021). Privacy and security issues in the future: A social media. Materials Today: Proceedings.

[15] Tian, W., Mao, J., Jiang, J., He, Z., Zhou, Z., Liu, J. (2018). Deeply understanding structure-based social network de-anonymization. Procedia Computer Science 129, 52–58.

[16] Wang, B., Jia, J., Zhang, L., Gong, N.Z. (2019). Structure-based Sybil detection in social networks via local rule-based propagation. IEEE Transactions on Network Science and Engineering 6(3), 523–537.

[17] Yang, D., Qu, B., Cudr'e-Mauroux, P. (2019). Privacy-preserving social media data publishing for personalized ranking-based recommendation. IEEE Transactions on Knowledge and Data Engineering 31(3), 507–520.

[18] Zhang, J., Sun, J., Zhang, R., Zhang, Y., Hu, X. (2018). Privacy-preserving social media data outsourcing. In: IEEE INFOCOM2018 - IEEE Conference on Computer Communications, 1106–1114.