Secure Analysis of Social Media Data

Hareesha Katiganere Siddaramappa
Department of Computer Application
Manipal Institute of Technology
Manipal Aacademy of Higher Education
Manipal, Karnataka, India
hareesh.ks@manipal.edu

Sumana Maradithaya
Department of ISE
MSRIT,Bangalore
Karnataka, India
sumana.m@msrit.edu

Sampath Kumar
Dept. of ECE,
Manipal Institute of Technology
Manipal Aacademy of Higher Education
Manipal, Karnataka, India
write2sk@gmail.com

***Abstract*—Confidentiality of the social media data during analysis is a major concern. Several real evidences show how the privacy and security of the data is compromised. One of the essential processes with social media data is to find the shortest paths between selected pair of nodes. This paper proposes a technique to modify the original data before analysis. The algorithm calculates shortest paths (data utility) between target nodes and then classifies edges into partially visited, all-visited and unvisited edges. Each category of edges is then perturbed using a dynamic variable value that is bound to satisfy specific constraints such that the shortest path as well as the shortest paths lengths, between the target node pairs remains the same. This paper proposes an approach to preserve the privacy of the weights and also generates an accurate length of the shortest path. It is also observed that the shortest path lengths between any target pairs of nodes are retained. The output is in the form of graphs, that shows that the proposed perturbation strategy perturbs the sensitive edge weights up to a maximum 72% , while keeping the difference in shortest path lengths minimum (up to 3%). It is hence demonstrated that along with preserving the sensitive information by perturbing the edge weights, the data utility is preserved i.e. the shortest path lengths are kept as near as potential to the original ones.**

Keywords—social networks, shortest path, privacy preserving, perturbation method.

I. INTRODUCTION

The proposed work emphasizes on sustaining the privacy of the edge weight in a graph. To ease the data analysis process, data owners might not want to perturb the shortest path length of a set of nodes but may not need to share the precise weight of every edge. A Perturbation Strategy is proposed which can retain the exact direct paths and tries to make corresponding lengths nearby to the original ones. This paper emphases on a privacy preservation procedure which is pragmatic on graphs to preserve data privacy and data utility. The data privacy is maintained with respect to the individual edge weights that is local information. Data Utility is for the shortest path, i.e., a path with a minimum sum of weights which is essentially a global property. While maintaining data utility, edge weights are perturbed as much as possible. The shortest paths and lengths approximate to the original ones as much as possible. As mentioned in [1] due to issues with high dimensionality and large scale of the data, traditional transformation techniques cannot be used to modify the original social network data. Another essential issue is in identifying what information in social networks is confidential and its relationship to personal privacy [2]. For instance, it is argued that associations in the form of weights when attached to edges are sensitive data and has to be preserved to avoid breach in privacy of the data in social networks. It is also difficult to mathematically define and manipulate data in social networks and quickly process such data to keep its privacy. Based on the above reasons, new theoretical foundations and corresponding technologies should be proposed to successfully and confidentially discover invaluable information in non-traditional data domains like social networks with a guarantee of privacy preservation within a satisfactory level. Recent works on privacy preservation in social networks [3, 4, 5] focus on de-identification procedure. These processes safeguard the privacy of persons while maintaining the patterns generated by interaction. These de-identification methods are frequently used when the individual's credentials are considered to be of top priority, example a customer's identity. But under several situations, the individual distinctiveness is not always measured to be trustworthy. In a weighted social network, the de-identification process without taking weight privacy into account is not enough to ease public privacy concern as node identifications are not considered as privacy in all cases. Also, some distinguishable weights can be used to reveal certain sensitive relationships if the weights are not modified in a weighted privacy preserving social network. Basic workflow of the model is shown in figure 2. EIES (Electronic Information Exchange System) [12] data is used in the project which is 32 x 32 matrix. A sample of the dataset is as shown in figure 1. After feeding the dataset to the model, preprocessing techniques are applied to convert the data into symmetric matrix according to the algorithm's requirement. Greedy perturbation algorithm is used in the project which helps in converting the symmetric matrix to get a perturbed adjacency weighted matrix. Further to analyze and visualize the results of output, graphs are used. The final output of this project will be graphs which give the estimation of the percentage of privacy that has been preserved along with percentage of change in shortest path lengths after using greedy perturbation algorithm.

EIES holds interactions between 32 researchers who communicated with each other through email. The tabulated data is shown in figure 1. Numbers 0 to 4 in figure 1 means the following: 0 means that no interactions have occurred, 1 means that the researcher had heard about the other but has not met him/her, 2 means that they have met each other, 3 means a researcher who is also a friend and frequently visited, 4 means a close personal friend with sufficient interactions, whereas 7 and 9 mean missing interactions.

II. LITERATURE SURVEY

According to L. Liu, J. Wang [6], there has been a large volume of privacy-preserving data extraction reviews in the literature. Many researchers attempt to develop methods to maintain data applications by not disclosing the original data and to produce data analytical results that are as close as possible to that of original data.

9	1	4	5	4	5	6	2	4	4
1	9	6	4	8	9	4	5	4	4
4	4	5	4	4	4	4	5	1	2
7	4	4	5	4	6	1	2	1	1
9	1	3	3	4	4	4	4	1	2
3	4	4	5	5	5	2	2	6	6
4	4	5	1	3	5	4	4	6	4
5	5	5	5	5	5	5	4	3	2
9	7	2	3	4	4	4	4	5	4
2	1	3	7	7	4	5	5	4	5

Fig. 1. Sample EIES data set comprising of acquaintanceship values between researchers

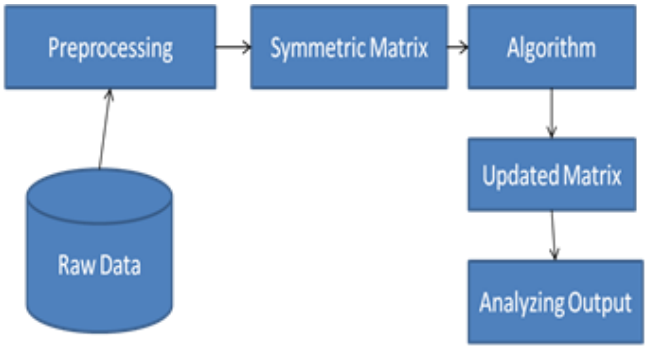

Fig. 2. Flow Diagram of the Proposed Work

Among those techniques [7], privacy preserving data mining can be broadly classified into two main categories. One category performs data mining by modifying the values of the distributed datasets and models them without knowing the exact values. Methods in the different category perturb the values of the dataset to preserve privacy of the data attributes. Perturbation techniques are divided into two subcategories, data addition and data multiplication, both of which are easy to implement but practically useful. In [13, 15], a structure for privacy preserving social network publication using the concept of grouping and anonymization. [14] Uses approaches to find the nearesh shortest path rather than perturbation.

According to Olivera Grljević, Renata Mekovec[8], for the data additive perturbation strategy, although individual data items are distorted, the aggregate properties of the original data can be accurately maintained. These properties may facilitate data clustering and classification and finding association rules. Data multiplicative perturbation is also good for privacy-preserving data mining. This technique dramatically distorts the original data but maintains inter-data distances which are also effective for distance specific applications such as clustering and classification. The difference between the two perturbation strategies is that, in the former strategy, only the aggregate distribution properties are available for data mining and the individual data behavior is hidden, while in the latter case it can keep more data-specific properties such as distances which can facilitate more diverse data mining tasks.

Liu, Lian et.al [9] clearly mentions the importance in preserving privacy in social networks and the use of traditional privacy preserving data mining. Perturbation is an secure method used to provide sufficient privacy in social network data. Here we provide a brief survey on privacy preserving social networks. Much progress has been made in studying the properties of social networks, such as degree of a node, structure, interactions (type and number) and identifying the society of people involved. As data is not structurally represented in matrix form in social networks, traditional matrix-based algorithms cannot be used to preserve privacy as mentioned in [10].

Authors emphasize on protecting the privacy of social entities using identification through de-identification techniques. Zhou et al in [11] discusses a framework of inserting and eliminating unweighted edges in social networks. This avoids impersonators in recognizing based on the information collected about the neighborhood. Emphasis was made on the protection of social entity's identification via de-identification using k-anonymity and its variants. However, k-anonymity methods work well on centralized data. [16] Elaborates on the need of social network analysis for identifying patterns among a group of researchers.

III. PROPOSED WORK

A social network graph is represented as G consisting of edges and vertices. The collection of edges is denoted by E and V represents the list of vertices. P indicates the shortest path. $w_{i,\ j}$ indicates the original weight of an edge between nodes i and j, ps1, s2 represents shortest distance between s1 and s2 whereas $d_{s1,\ s2}$ represents the minimum path length between s1 and s2. Being applied the perturbation, $w^*_{i,j}$, $p^*_{s1,\ s2}$, $d^*_{s1,\ s2}$ represents the perturbed weight, shortest path and perturbed shortest path length.

It is essential to observe that the modified graph will have the same name of vertices and edges as that of the unperturbed graph. However, the weights are enhanced and vary from the original graph. The path between the edges remains the same when compared to the earlier graph. The edges can be broadly classified as non-betweenness edge, all between edge and partial betweenness edge. When none of the shortest paths pass through an edge, that edge is called non betweenness edge. When all the shortest paths traverse through an edge then that edge is called all betweenness edge. All other edges are called partial betweenness edges.

A. Data Preprocessing

The following transformation is performed to generate the symmetric matrix 'W' where $W_{i,\ j} = 9 - (E_{i,\ j} + E_{j,\ i})$, where

$E_{i,j}$ is a numerical value which represents the i^{th} researcher's original association to the j^{th} researcher.

B. Calculation of Shortest paths

Floyd Warshall algorithm is used to calculate the shortest path between targeted pairs of nodes (H) from real network, using the adjacency matrix. The edges obtained are stored in a Dictionary with their count of occurrences.

C. Classification of Edges

The edges so obtained are classified into:

- All Visited - decreasing its weight will not change all shortest paths in H but decrease the length of corresponding shortest paths.
- Partially Visited - increase or decrease weight by 't'
- Non-visited - increasing its weight will not change all shortest paths and lengths in H.

Figure 3 shows all the three categories of edges after classifying them into visited, non-visited and partially visited

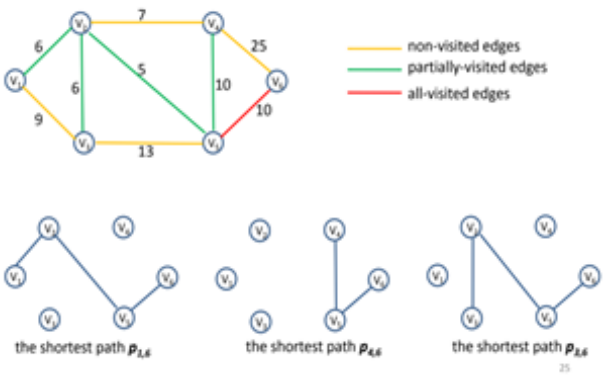

Fig. 3. Categories of edges

D. Improved Greedy Perturbation Algorithm

Notations Used

$w_{i,j}$ - The original weight of an edge between nodes i and j
$p_{s1,s2}$ - Shortest path between s1 and s2
$d_{s1,s2}$ - Shortest path length between s1 and s2.
$w^*_{i,j}$ - perturbed weight of edge between nodes i and j
$p^*_{s1,s2}$ – perturbed shortest path between s1 and s2
$d^*_{s1,s2}$ - perturbed shortest path length.

The weight of the partially visited edge Ei,j is enhanced by 't' where 't' is a value between 0 to minimum of the distance between the conditional shortest path($d_{+s1,s2}$) and distance between shortest path length from nodes s1 and s2($d_{s1,s2}$), for all shortest path from s1 to s2 ($p_{s1,s2}$). In the above case except the distance between the nodes are modified and becomes larger but the shortest paths remain the same. A graph of the conditional shortest paths are included in a graph G^+, which contains only the edges ei,j and ej,i and their corresponding weights ie ($w^+_{i,j}=w_{i,j}+ t$). For each node pair (s1, s2), $d_{s1,s2} \leq d^+_{s1,s2}$.

The weight of the partially visited edge Ei,j is reduced by 't' , where t is in between 0 and $\min\{d_{s1,i} + w_{i,j} , d_{j,s2} - d_{s1,s2}\}$,for all $p_{s1,s2}$. Here the distance between the nodes are decreased.

From the above-mentioned conditions for modifying distance, a practical greedy perturbation process is as described in Algorithm 1. The input that goes to this algorithm is the adjacency weight matrix obtained from the original graph. Using the Floyd Warshalls algorithm, the shortest path and their respective lengths are obtained. Based on the information obtained all the edges are classified as non-visit, visited and partially visited edges. The weights of the unvisited edges are perturbed by adding a random value. Similiarly, the weights of the visited edges are reduced using the same random value. An initial perturbed weight matrix P* is generated along with the perturbed shortest length matrix D*. Further, the weights of partially visited edges are customized. At first, partially visited edges are sorted based on the descending order of the number of shortest paths passing through that edge. All these are stored in a stack *Stack*. The edges are popped one by one and perturbed based on the conditions mentioned earlier. A popped-out edge will never be put back in the stack again. Hence, perturbation occurs only once for the partial betweenness edge. The matrix D* is recomputed and updated by Floyd-Warshall algorithm.

Input: Graph G representation: symmetric adjacency weight matrix W; Shortest Path Matrix D.
Output: The perturbed symmetric adjacency weight matrix D∗ of the corresponding perturbed graph G∗.
Procedure:
1. Calculate the shortest path and distance matrix from W to obtain P and D.
2. Assign D to D*.
3. Identify the non-visited and all visited edges ($E_{i,j}$) and modify the weights $W_{i,j} = W_{i,j} \pm t$. Select t based on the conditions. Update D*.
4. Compute the shortest paths passing through the partial visited edges. Push the edges into the stack *Stack* in decreasing order of shortest paths.
Repeat till Stack is empty
Pop edge on top of stack to get $e_{i,j}$
num1 = number of values where $d^*_{n1,n2}$ less than or equal to $d_{n1,n2}$, where n1 and n2 are nodes.
num2 = number of values where $d^*_{n1,n2}$ less than or equal to $d_{n1,n2}$, where n1 and n2 are nodes.
If (num1 › num2)
identify a value t within the range and update weight for the edge i,j.
$w^*_{i,j} = w_{i,j} + t$
else
obtain a value t given the range and upgrade weight for the edge i,j as follows $w^*_{i,j} = w_{i,j} - t$
end if
update the matrix D∗
end while.

IV. Results and Discussion

Table-I shows the difference between original edge weights (original costs) and perturbed weights (perturbed costs) for four iterations. The amount of preserving the privacy and the performance of the proposed perturbation algorithm is measured by mapping the edge weights and shortest path length. Figure 4 shows the improvement shown by the proposed model. The plot shows the percentage of perturbed shortest path lengths and weights following

perturbation. Also, on perturbation about 77% of the targeted pairs are preserved. The percentage of modified weight with length that descends within the x-axis difference to original ones is shown in figure 4.

TABLE I. COMPARISON OF ORIGINAL AND PERTURBED SHORTEST PATH COSTS IN VARIOUS ITERATIONS

Target Pairs	*Original Cost*	*Perturbed Cost 1st*	*Perturbed Cost 2nd*	*Perturbed Cost 3rd*	*Perturbed Cost 4th*
17:914	73	67	52	70	61
1:991	59	51	61	69	66
136:722	77	90	122	146	107
578:81	54	55	33	58	56
126:73	45	45	61	49	58
636:273	69	72	69	57	65
703:2	68	59	72	72	70
800:19	48	19	68	41	51
142:54	49	42	43	67	73
137:720	63	69	63	56	61
594:105	23	29	55	28	37
766:32	55	37	51	55	52
258:27	91	73	88	86	94
823:938	73	88	75	86	94
297:464	56	67	53	46	57
800:48	59	57	68	73	60
195:453	57	69	58	47	61
86:540	69	81	71	83	70

The plots show that the even after perturbation the modified shortest path length and the original shortest paths lengths are close enough. Hence the conclusions derived from this perturbed result would provide similar results. With reference to the Figure 5 the plot of original edge weights and the new perturbed edge weights which is obtained after the algorithm is applied. The Figure 6 shows the plot of original shortest path lengths and the new shortest path lengths obtained after the perturbation is applied on edge weights as per the algorithm.

With reference to Figure 7, the percentage change of edge weights on targeted edges. The graph shows the amount of privacy preserved. By perturbing the original weights of each edge between the targeted pairs we aim at preserving privacy. As shown in the Figure 8 the percentage change in shortest path lengths after the algorithm is applied. Around 80% of targeted pairs have their path lengths equal to the original ones. From this result we can conclude that the data utility has been preserved along with the data privacy.

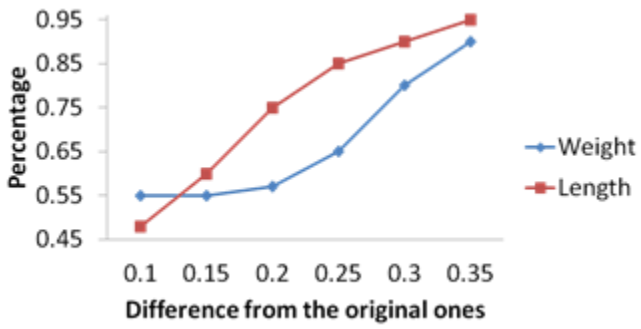

Fig. 4. Results with the improved greedy perturbation approach

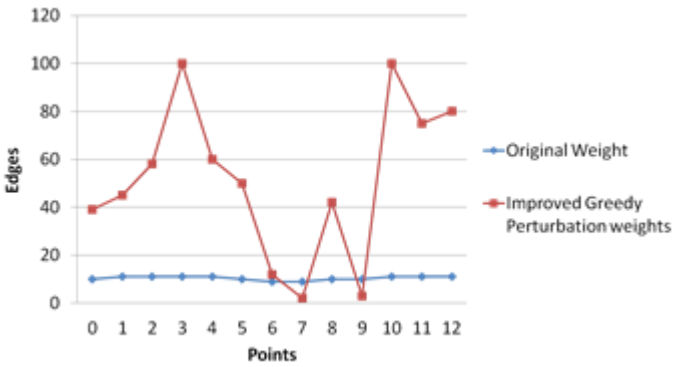

Fig. 5. The original edge weights vs perturbed edge weights of targeted nodes

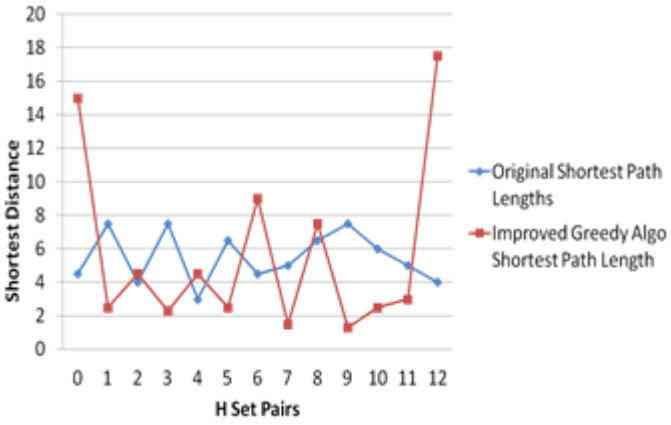

Fig. 6. Shows the plot of the original shortest path lengths and the perturbed shortest path lengths.

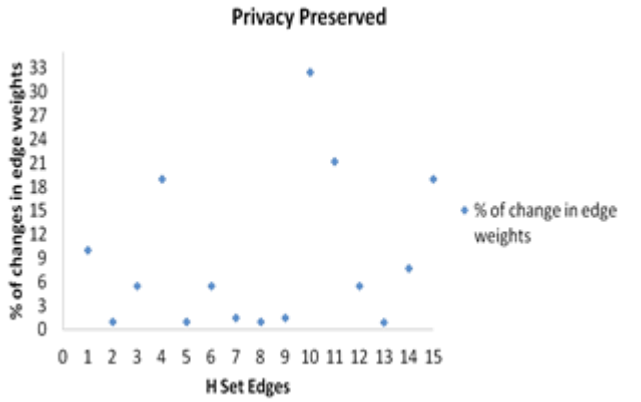

Fig. 7. Percentage of perturbation of edge weights

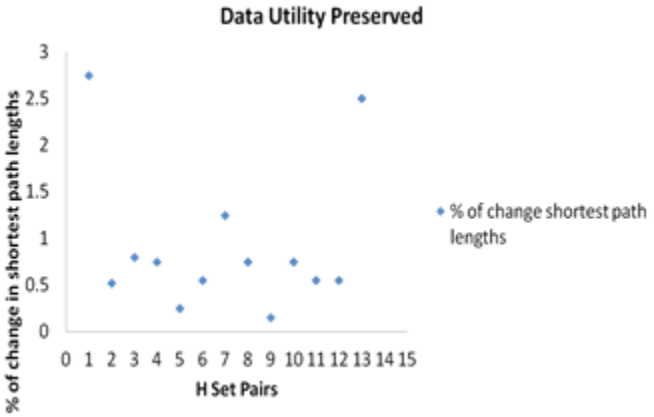

Fig. 8. Percentage of change in shortest path lengths between the target node pairs.

V. CONCLUSIONS

Social networks can be analyzed to discover various social issues like disease transmission, emotional contagion, and occupational mobility. With the advancements in technologies, people from various communities interact with each other regularly. This paper is motivated by the breach in the privacy concerns of the data while interacting. The privacy preserving framework allows data owners to maintain the confidentiality of the data while processing the necessary information. The proposed work protects the weights between the nodes that are considered susceptible but maintains the shortest paths and their path lengths. The experimental results demonstrate that the proposed perturbation strategy perturbs the sensitive edge weights upto a maximum of 72%, while keeping the difference in shortest path lengths minimum (upto 3%). This technique has various applications, for instance - in commercial data analysis field, this algorithm can help companies make better decisions such as choosing an optimal supply chain in the network and at the same time preserve sensitive information such as transaction expenses, bidding quotations etc . This algorithm is even useful in healthcare domain. It can provide better analysis of protein samples and DNA molecules to the biologists and medical researchers, while keeping the data of patients confidential.

REFERENCES

[1] Bonato A, Gleich DF, Kim M, Mitsche D, Prałat P, Tian Y, et al. (2014) "Dimensionality of Social Networks Using Motifs and Eigenvalues". PLOSOne journals.

[2] Zhang Z, Brij B.Gupta (2018) "Social media security and trustworthiness: Overview and new direction" Volume 86, September 2018, Pages 914-925, Future Generation Computer Systems.

[3] Leucio Antonio Cutillo, Refik Molva, Melek Onen(2011) , "Analysis of Privacy in Online Social Networks from the Graph Theory Perspective" , IEEE Globecom 2011 proceedings.

[4] Jiliang Tang, Huan Liu (2012) "Feature Selection with Linked Data in Social Media", SDM, 2012.

[5] A. Acquisti, and R. Gross. "Privacy Risks for Mining Online Social Networks". In NSF Symposium on Next Generation of Data Mining and Cyber-Enabled Discovery for Innovation (NGDM 2007), Baltimore, MD, October 2007.

[6] L. Liu, J. Wang, J. Liu, and J. Zhang, "Privacy preserving in social networks against sensitive edge disclosure," Tech. Rep. CMIDA-HiPSCCS 006-08, Department of Computer Science, University of Kentucky, 2008.

[7] Aggarwal, and C. C. Aggarwal. On the Design and Quantification of "Privacy Preserving Data Mining Algorithms".Madison, Wisconsin, June, 2002.

[8] Olivera Grljević, Renata Mekovec, Zita Bošnjak, "Privacy Preservation in Social Network Analysis", Croatia, September 19-21, 2012.

[9] Liu, Lian, "Privacy Preserving Data Mining For Numerical Matrices, Social Networks, and Big data". Theses and Dissertations--Computer Science, 2015.

[10] Zheleva and L. Getoor, "Preserving the privacy of sensitive relationships in graph data", in Proceedings of the 1st ACM SIGKDD International Workshop on Privacy, Security, and Trusting KDD, San Jose, California, pp. 153-171, Aug 2007.

[11] Zhou and J. Pei, "Preserving privacy in social networks against neighborhood attacks", in Proceedings of the 24th International Conference on Data Engineering (ICDE'08), Cancun, Mexico, pp. 506-515, April 2008.

[12] https://www.stats.ox.ac.uk/~snijders/siena/EIES_data.htm

[13] Tsan-sheng Hsu, Churn-Jung Liau ∗, Da-Wei Wang, "A logical framework for privacy-preserving social network publication", Journal of Applied Logic, Volume 12, Issue 2, June 2014, Pages 151-174.

[14] Nayan Mattani, J. Sharath Kumar, A. Prabakaran and N. Maheswari, "Privacy Preservation in Social Network Analysis using Edge Weight Perturbation ", Indian Journal of Science and Technology, Vol9(37), October 2016

[15] Maria E. Skarkala, Manolis Maragoudakis, Stefanos Gritzalis and Lilian Mitrou Hannu Toivonen and Pirjo Moen, "Privacy Preservation by k-Anonymization of Weighted Social Networks", 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.

[16] David Dietrich, Barry Heller, "Data Science &Big Data Analytics", 2015, Wiley Publications.