Preserving Privacy of Social Networks Data Against Mutual Friends and Degree Attacks

Amardeep Singh, Divya Bansal, Sanjeev Sofat
PEC University of Technology
India
amardeepsingh_26@yahoo.com

ABSTRACT: *Online social networks have become a predominant service on the web collecting huge amount of users' information. It is drastically revolutionizing the way people interact. Publishing data of social network users for researchers, academicians, advertising organizations etc. has raised many serious privacy implications. Lots of techniques have been proposed for preserving privacy of individuals handling different types of attack scenarios used by adversaries. In this paper, we address a new attack i.e. mutual friends attack model, in which an adversary can identify the victim nodes by using knowledge about the number of their mutual friends. An algorithm 'Optimized K-anonymization' has been devised that can deal with two types of attacks i.e. degree attacks and number of mutual friends attacks. The experimental results illustrate that our proposed algorithm can preserve the identification of individuals and subsequently maintain the utility of data.*

Classification: CCS → Security and Privacy → Human and Societal Aspects of Security and Privacy → Privacy Protections

Keywords: Privacy Preserving, Social Networks, Degree Attacks, Mutual Friends Attacks

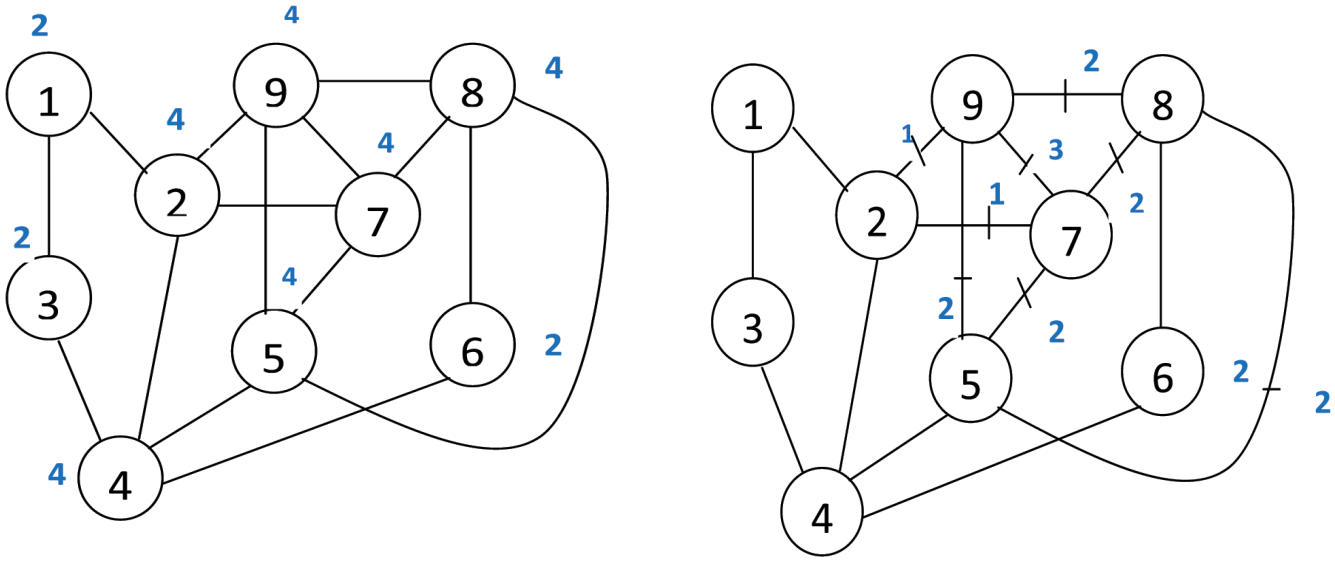
Received: 20 May 2018, Revised 19 June 2018, Accepted 3 July 2018

DOI: 10.6025/jio/2018/8/3/83-97

1. Introduction

Advancement of Internet has attracted billions of users to become a part of social networks such as Facebook, Twitter, Google+, WhatsApp etc. Consequently, these online applications are collecting personal information of users like their name, age, date of birth, location, school details, college details, job information, hobbies, mobile number etc. The implicit information contained in the social network data offers interesting problems for data mining and information extraction. Therefore, publishing of the data is essential from the perspective of research. Publishing of social networks data has led to the risk of leaking sensitive and confidential information of individuals and may cause many serious privacy threats to the individuals. This requires preserving the privacy of individuals before publication of social networks data. Privacy of Online social networks data has been of utmost concern in the recent years. To protect privacy of social network users, lots of work has been done by researchers (Zhou et al., 2008).

Sun *et al*. (Sun *et al*., 2013) proposed a novel attack called 'mutual friends attack'. In this attack, an adversary can identify a victim node in a graph by using the number of mutual friends of two end vertices of an edge. He further addressed the issue that more attacks are possible in a social network other than degree attacks as huge information of users is available on social networks (Sun et al., 2013). Figure 1 shows an example of mutual friend attack on K-anonymized graph (social network). Figure 1(a) shows a 2-anonymized graph with the degree of each node. In this anonymized graph of Figure 1(a), an attacker cannot identify the victim node by possessing the background knowledge about the degree of nodes, as all the nodes are indistinguishable from other (K-1) nodes. Figure 1(b) shows the 'number of mutual friends, of different edges in an anonymized graph of Figure 1(a). If an adversary knows the number of mutual friends of victim node as '3', then he can easily identify the edge (7,9) connecting nodes (7) and (9), as only this edge has „number of mutual friends as '3' as shown in Figure 1(b).

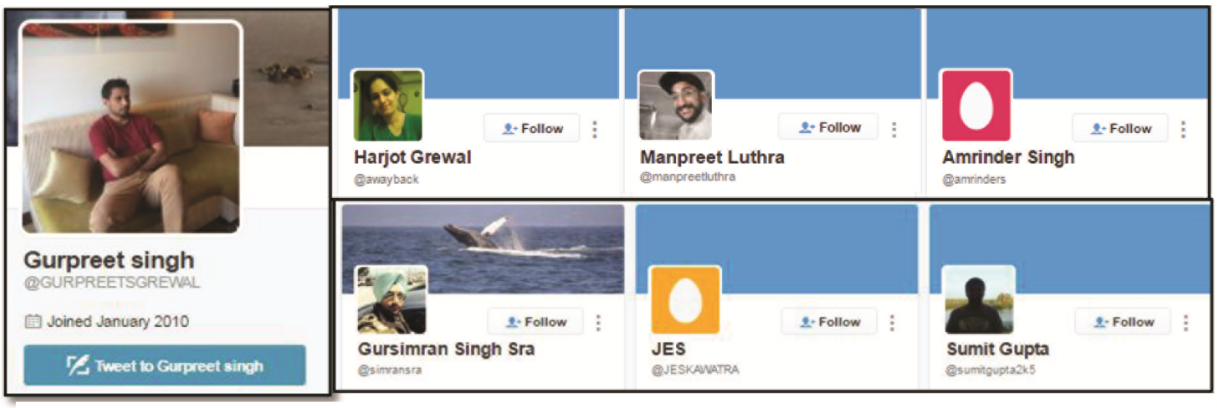


(a) 2-anonymized graph with degrees of nodes (b) Mutual friends attack on 2-anonymized graph

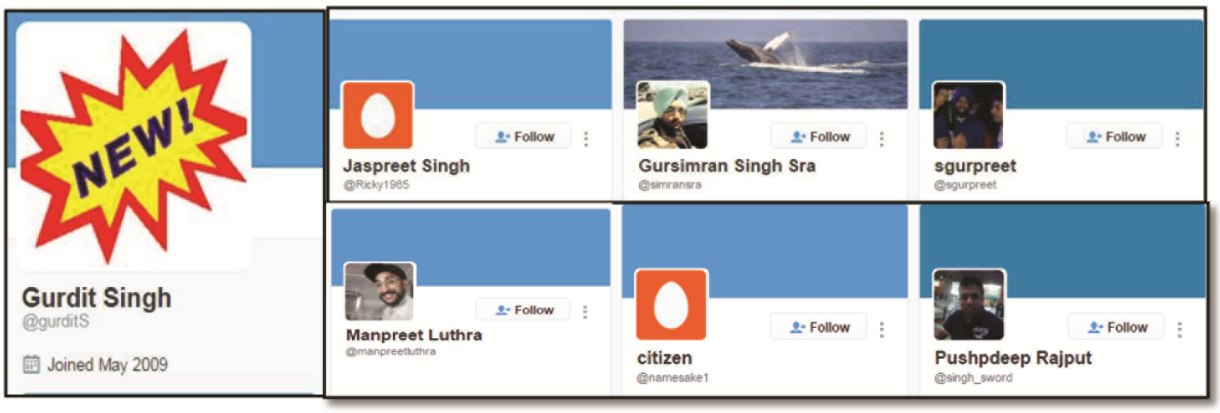
Figure 1. Example of mutual friends attack on K-anonymized social network

This example shows that even after applying K-anonymity on social networks, an adversary can easily identify victim nodes by 'mutual friends attack'. And consequently, sensitive information associated with vertex nodes can be revealed to the attackers.

In social networks such as Twitter and Facebook, the information of mutual friends of users can be easily obtained. Figure 2 shows an example from Twitter social network, where followings (also called friends, whom the user is following) of a user can be easily seen. Hence, one can easily figure out the number of mutual friends of two Twitter users.



a) Friends of Twitter user with screen name "@GURPREETSGREWAL" (Friends of Twitter user "@GURPREETSGREWAL")



(b) Friends of Twitter user with screen name "@gurditS" (Friends of Twitter user "@gurditS")

Figure 2. Friends list of users on Twitter

In Figure 2(a), friends of user with screen name '@GRPREETSGREWAL' (Friends of Twitter user "@GURPREETSGREWAL") are shown and Figure 2(b) shows the friends of user with screen name '@gurditS' (Friends of Twitter user "@gurditS"). Information about the number of mutual friends of both the users can be obtained by intersecting the friends list of two usersual friends of two no. From Figure 2, the number of mutual friends of two Twitter users is '2'.

An *'Optimized K-anonymization algorithm'* has been proposed to deal with 'mutual friends attacks' and 'degree attacks'. This algorithm is based upon accomplishing the principles of K-anonymity by using mutual friends of two nodes as noisy nodes.

1.1 Our Contributions

Our main contributions in this paper can be summarized as follows.

1. The proposed *'Optimized K-anonymization algorithm'* is capable of handing degree attacks as well as mutual friends attacks.

2. The proposed algorithm is using mutual friends of two nodes as noisy nodes to achieve the property of K-anonymity. This helps in preserving the original graph structure and hence achieves better utility.

3. The experimental results on real-time social networks show better performance of our proposed algorithm measured in terms of average path length and information loss.

The rest of the paper is organized as follows. Firstly, in section 2, related work in the domain has been discussed. The problem formulated based upon literature survey has been presented in section 3. Proposed algorithm has been presented in section 4. Experimental results have been shown in section 5 followed by conclusions in section 6.

2. Related Work

Several published academic works have proposed techniques for preserving privacy of micro-data and social networks data. The domain of the literature review is the techniques that ensure privacy of data. It serves the purpose to understand the implications of privacy preservation. In this section, the techniques based on K-anonymity, L-diversity, clustering and combined approach of K-anonymity, L-diversity in social networks has been discussed.

The concept of K-degree anonymity (Sweeney, 2002) was introduced to protect identification of individuals from an adversary who possess the knowledge about the degrees of nodes. It states that in an anonymized network, each node should possess the same degree as other (K – 1) other nodes. Zhou and Pei (Zhou and Jian, 2008) utilized this property to protect individuals from background knowledge attack. A graph is said to satisfy K-neighborhood anonymity if neighbourhood of every node in the network is similar to the neighborhood of at least (K – 1) other nodes. Wu et al. (Wu et al., 2008) formally named this property 'K-neighbourhood anonymity. K-Candidate anonymity (Hay et al., 2009), targets the structural anonymity of users in the presence

of subgraph structure of the network. Hay et al. defined three types of structural queries - vertex refinement queries, subgraph queries and hub fingerprint queries (Hay et al., 2008) (Hay et al., 2009). Zou et al. (Zou et al., 2009) proposed a technique of K-automorphism anonymity to protect against adversary who possesses the information about the subgraph of a victim node.

In another important piece of work, Liu and Terzi (Liu and Terzi, 2008) anonymized the network to satisfy K-degree anonymity by using the concept of edge modification techniques to meet desired constraints. The objective of the algorithm is to achieve anonymity through minimizing the additions and deletions of edges. In a similar type of work, Zou et al. (Zou et al., 2009) proposed an edge modification algorithm. Another variant of edge modification is 'anonymization by randomization'. This approach modifies the graph structure by deleting and adding edges randomly but preserving the total count of edges. But later on, Hay et al. (Hay et al., 2009) proved that edge modification fails to preserve significant graph metrics of real-world networks. Ying and Wu (Ying and Wu, 2008) proposed spectrum-preserving randomization to retain the utility of social networks. The spectral properties of a graph are eigenvalues of the graph s adjacency matrix. Preserving these properties directs the selection of random edges for addition and deletion. However, the privacy preservation results of this approach are uncertain.

Two significant works have presented technqiues for reconstructing randomized networks (Vuokko and Terzi, 2010) (Wu et al., 2010) . Wu et al. (Wu et al., 2010) applied a low rank approximation technqiue to a randomized network structure in order to obtain effective topological features. Vuokko and Terzi (Vuokko and Terzi, 2010) reconstructed networks by applying randomization both to the structure and attributes of the nodes. But in both the works, the outcome of reconstruction techniques on privacy preservation has not been evaluated.

In another significant work, Hay et al. (Hay et al., 2009) proposed an efficient technique that allows the publication of degree distribution of the nodes while preserving differential privacy. Further to produce more accurate results, a post processing step has been applied on the differentially private output. The experimental analysis on synthetic and real-world networks presents that the effectiveness of the approach results in low bias and variance.

Zhou and Jian (Zhou and Jian, 2008) showed that achieving k-neighborhood anonymity is a NP-hard problem and further proposed a greedy edge modification and label generalization approach. The main objective of the algorithm is same as Liu and Terzi (Liu and Terzi, 2008) i.e. to minimize the number of edge additions. Zheleva and Getoor (Zheleva and Getoor, 2008) assumed that an adversary utilizes a precise statistical model for identifying sensitive relationships in the presence of attributes of vertices and edges in the original graph. They proposed a two step approach. First of all, nodes are considered as records in a tabular form, and the attributes of records are anonymized using one of privacy definitions to preserve the privacy of users. Then, the structure of the network is preserved by keeping aggregate information about the structure of network between the equivalence classes. Network generalization approach has been further used by Campan and Truta (CampanA, TrutaTM) to anonymize a social network. Greedy approach has been used to optimize the utility of the network by using the structural and attributes information simultaneously rather than following a two-step approach. They introduced structural information loss as a measure of utility.

K-anonymity can prevent structure attacks as seen earlier in this section but it fails to avoid the leakage of sensitive attributes. The deficiency of K-anonymity can be conquered by L-diversity. Machanavajjhala et al. (Machanavajjhala et al., 2007) introduced the notion of L-diversity which states that the distribution of a sensitive attribute in each equivalence class has at least L well represented values. Panda et al. (Prasad et al., 2010) used the notion of L-diversity to preserve the privacy in social networks data. The performance of the approach has been analyzed in terms of utility of social networks. In another piece of work, Li et al. (Li et al., 2011) proposed two graph modification based algorithms to achieve L-diversity anonymization.

In another significant work, Kavianpour et al. (Kavianpour et al., 2011) introduced an integrted approach using combined strength of K-anonymity and L-diversity algorithms. The approach has been evaluated by using information loss as a parameter to check the effectiveness of their proposed work. In another piece of work, Tripathy et al. (Tripathy et al., 2012) presented an algorithm using the notion of K-anonymity and L-diversity based upon variants of multi-sensitive attributes. The proposed work protects users against neighbourhood attacks. One observed drawback of the proposed algorithm is its complexity. This technique requires improvements before evaluating on large datasets. Yuan et al. (Yuan et al., 2013) presented a novel concept of noisy nodes for the preservation of identity and sensitive attributes of vertices in a social network. Experiments have been conducted using social network datasets of ARNET, CORA and DBLP. Evaluation of the approach has been carried out using perormance metrics auch as APL (Average Path Length), ACSPL (Average Change in Sensitive Path Length), RRTI (Remaining Ratio of Top Influential People) and percentage of noise nodes etc. They introduced the notion of noisy nodes for preserving

privacy in social networks. Our proposed technique also uses the concept of noisy nodes to achive the property of K-anonymity but the differnce is that our algorithm is using mutual friends of nodes as noisy nodes.

Clustering is another technique to achieve anonymization by grouping vertices into a super vertex (Bhagat et al., 2009) (Cormode et al., 2008) (Tassa et al., 2013). One major drawback of this approach is the hiding of vertex and its relationship with other vertices. This results in changing the original structure of the network and makes the released data unusable.

From the comprehensive literature, it has been observed that a lot of work has been carried out to preserve privacy of social network users. But still some gaps have been observed as mentioned here. It has been observed that a new attack 'Mutual Friends attack' addressed by Sun et al. (Sun et al., 2013) has not been investigaed much in literature. It has been anticipated that more attention needs to be given to retain the utility of data. Developed techniques need to be validated on real-time social networks such as Facebook, Twitter, Google+ etc.

These gaps have been addressed in the present work as discussed in the subsequent sections of the paper.

3. Problem Definition

In this paper, a social network is presented as a 3-tuple undirected graph $G(V, E, L)$, where V is the set of nodes or vertices showing individuals, E is the set of edges showing the connection between vertices, and L is the set of labels assigned to the edges such as their degree, relationship, number of mutual friends etc. as shown in Figure 3.

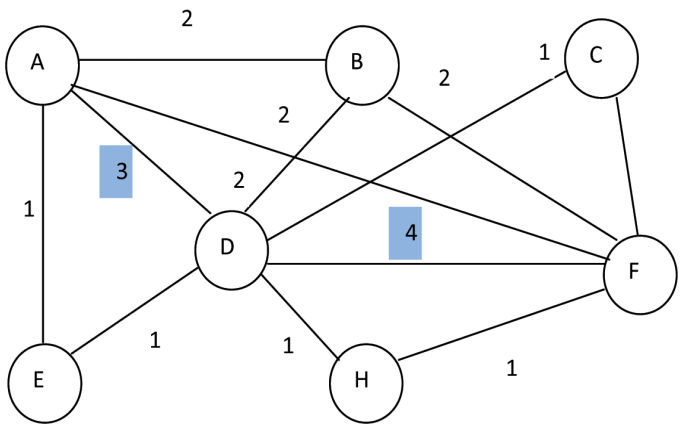


Figure 3. Original Graph

Definition 1: Number of Mutual Friends of an Edge

Suppose vertices $v1$ and $v2$ are connected to each other via an edge e1 in graph G i.e. $e_1=(v1,v2)$ and $v1, v2 \in V$ and $e1 \in E$. The number of mutual friends of an edge $e1$ is the number of nodes related to both vertices $v1$ and $v2$.

Let MF represents the sequence of mutual friends and E is the list of edges in graph G. Taking an example of Figure 3, set of number of mutual friends is $MF = \{2, 3, 1, 2, 2, 2, 1, 4, 1, 1, 1\}$ and list of corresponding edges is $E = \{(A, B), (A, D), (A, E), (A, F), (B, D), (B, F), (C, D), (D, F), (D,E), (D, H), (F, H)\}$.

Alternatively, the number of mutual friends of an edge $e \in E$ can also be measured by counting the number of triangles which contain edge e. For example, in Figure 3, number of mutual friends of vertices (A) and (D) is '3'. This can be obtained from the number of triangles containing edge (A, D) i.e. (A, D, B) , (A, D, E), (A, D, F).

Definition 2: Mutual Friend Attack

In mutual friend attack, an adversary possesses the information about the number of mutual friends MF_e of the victim edge $e1 \in E$ and uses this information to reach the victim node.

In case, MF_e is unique in graph G, then adversary can easily reach the target node by finding the corresponding edges in the list E. In Figure 3, if an adversary knows that MF_e of the victim node e is 4, then he can identify the target edge as (D, F) as there is only one edge in E with MF_{DF} as 4.

Definition 3: K-anonymity Principle
Graph G is said to satisfy the property of K-anonymity, if there exists K-1 other nodes similar to K in terms of properties such as degree, subgraphs, neighbourhood etc.

Problem Definition
For a graph G (V, E, L), the problem is to anonymize the graph such that degree attacks and mutual friend attacks can be handled by preserving utility of original graph.

Taking an example of Figure 3, both edges AD and DF can be identified by the adversary using the number of mutual friends and degree of the vertices A and D. Hence, addition/deletion of edges/vertices needs to be done to anonymize the graph by changing the original graph as less as possible.

4. Proposed Optimized K-Anonymization Approach

In this section, we present our proposed optimized anonymization algorithm to preserve the privacy of social network users while handling degree and mutual friend attacks.

To make the nodes indistinguishable from other (K-1) nodes, firstly all the nodes in the graph are divided into three groups based upon the average degree of nodes. First group will contain nodes with degree less than the average degree, second group will contain nodes with degree greater than average degree and third group will contain nodes with degree equal to the average degree. Next step is to make degree of all the nodes in three subgroups equal. For this, threshold value of degree i.e. average degree of all nodes in three subgroups is calculated. Degree of all the nodes in each subgroup is made equal by adding/deleting the noisy nodes. Mutual friends of nodes are taken as noisy nodes with the aim to preserve the utility of original data. Details of our proposed technique are given in Algorithm 1.

Optimized K-anonymization Algorithm

Algorithm 1: Preserving Privacy by handling Hegree and Mutual Friends Attacks

```
Input: Original network graph G (V,E,L)

Output: Anonymized network graph G (V,E,L) statisfying K-anonymity principle and
protecting against degree and mutual friends attacks

1. Upload data of edges and vertices.

2. Calculate degree of each vertex.

3. Calculate friend_list of each vertex v in graph G. Friend_list is the list of
verices conncted to each vertex.

4. Calculate number of mutual friends MF of edges in graph G.

5. Calculate mean degree of all the vertices. And consider this degree as a threshold
degree. threshold_social_network = (Min_degree+Max_degree)/2 //threshold value

6. Form 3 clusters (groups) based upon degree of vertices as follows:

      a) G1 = {Nodes with degree<threshold_social_network}

      b) G2 = {Nodes with degree>threshold_social_network}

      c) G3 = {Nodes with degree==threshold_social_network}
```

```
7. Form subgroups within three groups as follows:
      a) Calculate threshold_group = (Sum of degree of all vertices in a group)/No.
            of vertices in a group
      b) Form groups within groups as
            i) G11= {Nodes with degree< threshold_group}
            ii) G21= {Nodes with degree> threshold_group}
            iii) G31= {Nodes with degree== threshold_group}
8. Addition of vertices to increase degree in a subgroup
      a) Find vertices in a particular subgroup with degree > threshold_group
      b) Find Common vertices between the current node and nodes with degree
      >threshold_group using friend_list of vertices
      c) If found, then add this vertex to the current vertex with a different value
      of other attributes
      d) Otherwise add a random vertex with random values of attributes
Repeat step 8 till degree of all the vertices in a subgroup becomes equal.
9. Deletion of vertices to decrease degree in a subgroup
      a) Find duplicate vertices in the friend_list of two vertices under
            consideration
      b) If found, then delete duplicate vertices
      c) Otherwise delete a vertex randomly
Repeat step 9 till degree of all the vertices in a subgroup becomes equal.
10. Update the list of number of mutual friends of edges in graph G.
```

Our proposed Algorithm 1 has been applied on social network of Figure 3 and results obtained after each step is shown below:

After Step 1, 2, 3

This step returns the degree and friend_list of all the nodes in the graph as given in Table 1.

Vertex	Degree	Edge List / Friend_List
A	4	B, D, F, H
B	3	A, D, H
C	2	D, H
D	6	A, B, C, F, H, J
E	2	A, D
F	5	A, B, C, D, J
H	2	D, H

Table 1. Degree and Friend_list of nodes of Figure 3

After Step 4
This step gives the 'number of mutual friends (MF)' of edges in the graph as shown in Table 2.

Edge	Number of mutual friends (MF)
AB	2
AD	3
AE	1
AF	2
BD	2
BF	2
CD	1
DF	4
DE	1
DH	1
FH	1

Table 2. Edges and their number of mutual friends of Figure 3

Here, 'degree attack' is possible if the adversary knows that the victim node is connected to 6 other nodes. And „mutual friends attack is possible if adversary has the information that the number of mutual friends of victim node is 3. Using this background knowledge, an adversary will identify the victim node as *D* because edge (A, D) has 3 mutual friends and D is the only node that has degree 6 in graph G of Figure 3.

After Step 5
This step calculates the average degree for all the nodes in graph.

Here, threshold_social_network = 3.4

$$\sim 3$$

After Step 6
This step formulates three groups based upon the average degree of step 5.

Based upon the degree, following 3 groups are formed:

$G1 = \{C, E, H\}$ with corresponding degrees = {2, 2, 2}

$G2 = \{B\}$ with degree = {3}

$G3 = \{A, D, F\}$ with corresponding degrees = {4, 6, 5}

After Step 7
This step forms subgroups within three main groups as follows:

a) Threshold_degree within $G1 = 2$.

All the nodes have same degree and hence no sub-grouping is done.

b) Threshold_degree within $G2 = 3$

There is only one node with degree 3. Hence no further grouping is done.

c) Threshold_degree within $G3 = 5$

Here, three subgroups are formed as follows:

$G11 = \{A\}$, with degree < 5

$G12 = \{D\}$, with degree > 5

$G13 = \{F\}$, with degree $= 5$

After Step 8

This step increases the degree of nodes to equalize the degree to the threshold value of the subgroup.

In group $G3$, vertex $\{A\}$ has degree '4' and vertex $\{D\}$ has degree '6' . Degree of vertex $\{A\}$ is to be increased by '1' and degree of vertex $\{D\}$ is to be decreased by '1' .

First, the addition of edges is performed to increase the degree of nodes. From Step 1, the edge list of node $\{A\}$ and $\{D\}$ is checked. And common vertices of $\{A\}$ and $\{D\}$ are obtained as $\{B, F, H\}$. Considering first common vertex $\{B\}$, remove $\{B\}$ from vertex $\{D\}$ and add it to $\{A\}$ as $\{B\}$ with different attributes. (This is based upon the concept that if a user X is connected to user Y on Facebook in his school friend group and Y is also connected to X through his organization group. This way, both X and Y share two types of relationships with different relationship properties although users are same).

Step 9 is skipped as all the nodes have degree equal to threshold degree.

After Step 10

This step updates the number of mutual friends of all the edges in the graph.

After step 10, $MF_{AD} = 2$, $MF_{DF} = 3$

This example shows that original graph G is K-anonymized as G with different values of MF and degrees. Thus, our proposed algorithm is capable of handling 'degree' and 'mutual frinds' attacks with minimum changes in the original graph by preserving the utility of data. This has been shown in the subsequent sections by taking various performance metrics.

After applying our proposed '*Optimized K-anonymization algorithm',* there is no vertex with unique degree and *MFs*. And if mutual friends are unique, they are different from the original values. This way, the algorithm ensures that an adversary cannot identify the victim node.

5. Experimental Results

In this section, we present the experiments results conducted on data of social networks to evaluate the performance of our '*Optimized K-anonymization algorithm'*. Utility of the anonymized networks has been evaluated by computing average path length, and information loss.

5.1 Datasets

The experiments have been conducted on two datasets: CORA and Twitter. Both the datasets have been preprocessed to be presented in the form of linear tabular structure. All the attributes such as their relationship, weightage of relationship etc. are presented in numeric format.

CORA

This dataset is a collection of papers on Computer Science (Yuan et al., 2013). We have extracted seven different classes of papers on the subject 'machine learning'. If two authors have co-authored a paper then there is a connection between them. There are 2708 papers taken as vertices and 10557 connections taken as edges.

In Table 3, 'ID' represents an author connected to other author with „ID stored in column 3. Name in column 2 presents the name

of author whose 'ID' is saved in column 1. 'Weight' is the importance (weightage) assigned to the relationship of two authors (depending upon the number of papers co-authored) and 'Label ID' presents the one of the seven categories of the papers from machine learning discipline.

ID	Name	Edge ID	Weight	Label ID

Table 3. Attributes in CORA Dataset

Twitter

Data of Twitter users has been extracted using NodeXL (NodeXL). NodeXL is a data mining and analysis tool. We have crawled data of 66 users with 170 connections. Users are connected to each other through relationships such as 'Tweet' 'Mentions', 'Replies to', and 'Follows'. Entire data has been converted into numeric format as per our requirement. Five attributes have been extracted from the collected data and stored in a tabular form as shown in Table 4.

Vertex1	Name	Vertex2	Relationship Type	Location

Table 4. Format of Twitter Dataset

In Table 4, 'Vertex 1' is the source node having 'Name' stored in column 2, and connected to 'Vertex 2' with 'Relationship type'shown in column 4. 'Location' is the attribute presenting the location of 'Vertex 1'.

5.2 Performance Metrics

To measure the performance of our proposed anonymization algorithm in terms of utility, following metrics have been used:

5.2.1 Average Path Length (APL)

Average path length (APL) is defined as average of the shortest distance between all the vertices in a graph. Distance between nodes in a graph can be measured by various functions implemented in Matlab like Hamming distance, Euclidean Distance, and Mahalanobis distance etc. In the present work, Mahalanobis distance function has been used to calculate the value of average path length.

Mahalanobis distance (De Maesschalck et al., 2000) is calculated using equation (1).

$$\text{Distance} = \sqrt{((y - \bar{Y})^T S^{-1} (y - \bar{Y}))} \qquad (1)$$

Where, y = each of the N nodes

$Y = N X d$ matrix

d = Dimension of the data

S^{-1} = Covariance matrix

$\bar{Y}$ = Mean value of y

T = Transpose of data

This function calculates the distance of each observation y in Y to the mean of all components of the distribution. Calculated value is the distance of each y observations from the mean value of component Y. Minimum changes in values of APL after anonymization shows the effectiveness of the proposed algorithm.

5.2.2 Information Loss

The goal of our proposed 'optimized K-anonymization' algorithm is to preserve the utility of the anonymized social network by handling degree and mutual friend attacks. Information loss is an important metric for measuring utility of graphs. Information loss calculates the loss of information by measuring the change in the structure of graph after anonymization (Machanavajjhala et al., 2007). To calculate information loss, equation (2) has been used as given in literature (Machanavajjhala et al., 2007) (Ford et al., 2009).

$$Info\_loss = (x1-3)*((y1/(y1*(M-N)))+(y/y1)) \quad (2)$$

where

$[y, x] = size(data1)$;

$[y1, x1] = size(data2)$;

$data1$ = *original data*;

$data2$ = *anonymized data*

Values of M and N are calculated using the function as shown in Figure 4.

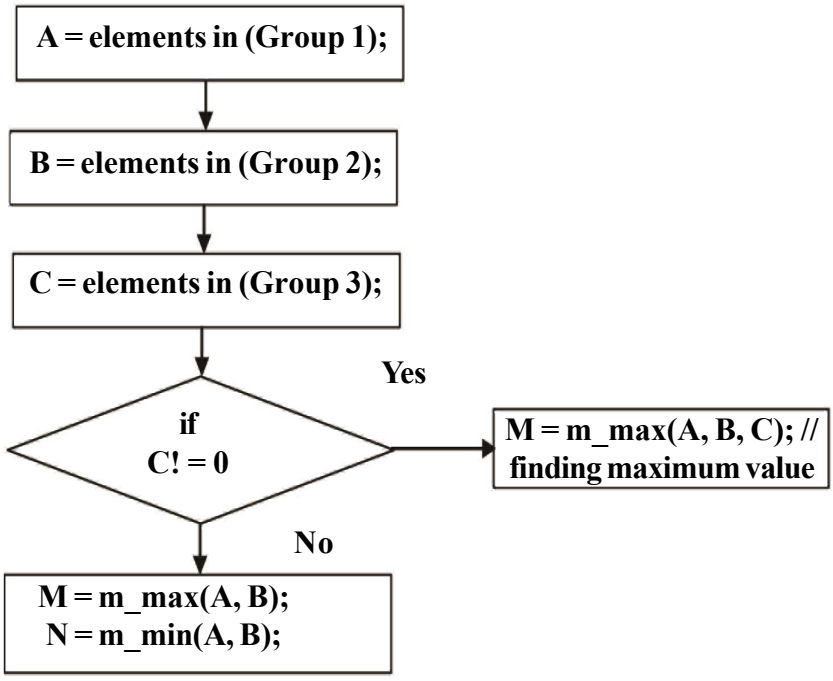


Figure 4. Calculating max and min values in Information loss

These two metrics have been used in literature (Machanavajjhala et al., 2007) (Yuan et al., 2013) (Ford et al., 2009) for evaluating the performance of anonymization algorithms. These metrics evaluate the utility of social networks data. For both the metrics, the value of original social network has been compared with the value of anonymized social network. Closer value of APL of anonymized social network to the APL of original social network shows that utility of original network is attained and better utility indicates better performance. And less value of information loss after anonymization indicates better utility of the graph.

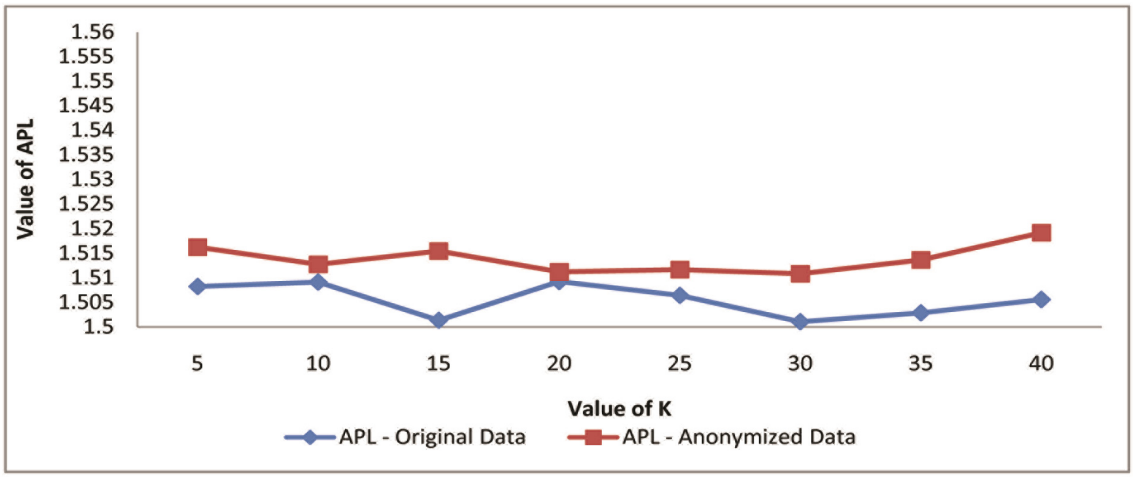
5.3 Evaluation of Proposed Algorithm

In this section, the evaluation results of our proposed 'optimized K-anonymization' algorithm have been discussed.

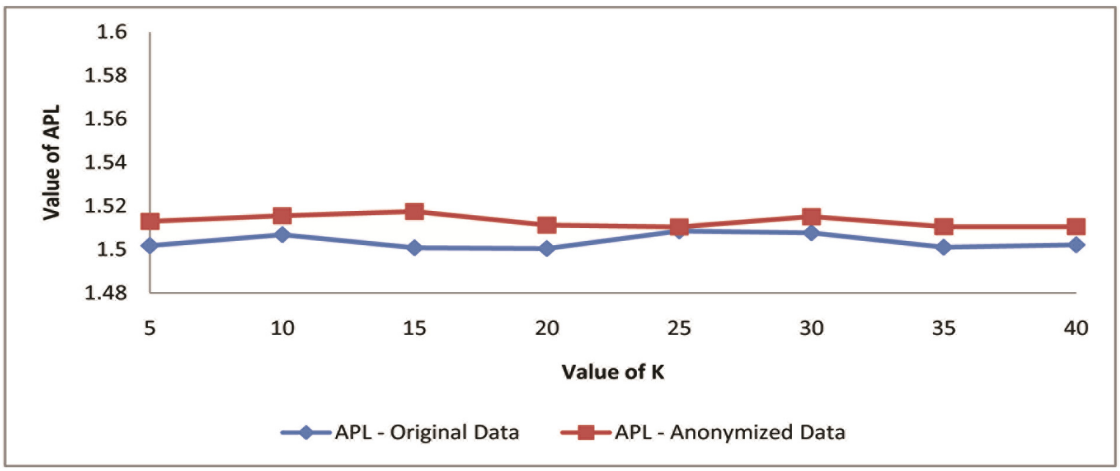
5.3.1 Average Path Length (APL)

We first compare average path length (APL) of the anonymized social network with *APL* of original graph. Results of comparison are shown in Figure 5. It is clear from Figure 5 that the difference of APL values of anonymized graph differs slightly from the APL values of original graph for both the datasets. The largest difference value is 0.014 when K is equal to 15 in the CORA dataset. And in Twitter dataset, the largest difference is 0.016703 when K is set to 15.

Figure 5 shows the values of APL for both datasets with different values of K ranging from 5 till 40. The values of APL before and after anonymization are also shown in Table 5 and Table 6 for CORA and Twitter respectively.



(a) CORA



(b) Twitter

Figure 5. Values of APL before and after anonymization

5.3.2 Information Loss

Figure 6 shows the information loss obtained for both the datasets using our proposed algorithm. Generalized information loss has been calculated by measuring the change in the structure of graph after anonymization. Less value of information loss after anonymization presents better performance of the approach.

Figure 6 shows that loss of information after anonymization in CORA dataset is 1.5329% whereas in Twitter it is very less i.e. 0.14082%.

From the evaluation results of APL and information loss, it is clear that our *'optimized K-anonymization'* algorithm is capable of preserving the utility of original graph effectively by preserving 'degree' and 'mutual friends attacks'.

6. Conclusions

The information shared in different social networks are under high risk of attacks by various adversaries, in other words, privacy of its users could be easily breached by this information. For a service provider, such as Facebook, Twitter and Linkedin, publishing a privacy preserving social networks data has become an important issue. It thus becomes essential to preserve

K	APL - Original Data	APL - Anonymized Data	Difference in APL
5	1.508247	1.51627	0.008023
10	1.509158	1.512766	0.003608
15	1.50137	1.515471	0.014102
20	1.509234	1.51121	0.001977
25	1.506424	1.511693	0.005269
30	1.501075	1.510843	0.009768
35	1.502885	1.513649	0.010764
40	1.505569	1.519204	0.013635

Table 5. Values of APL before and after anonymization for CORA dataset

K	APL - Original Data	APL - Anonymized Data	Difference in APL
5	1.501697231	1.512937566	0.01124
10	1.506810571	1.515550145	0.00874
15	1.500783148	1.517486539	0.016703
20	1.500397476	1.511198006	0.010801
25	1.50851617	1.510378058	0.001862
30	1.507639254	1.515077748	0.007438
35	1.501077849	1.510497327	0.009419
40	1.502184849	1.51053206	0.008347

Table 6. Values of APL before and after anonymization for Twitter dataset

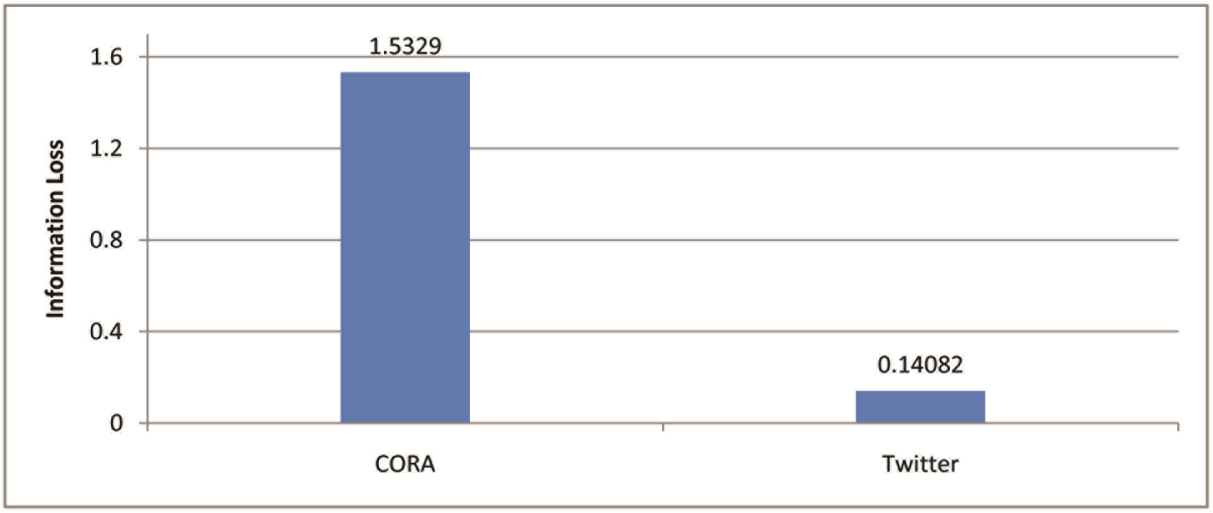


Figure 6. Information Loss for both datasets

users' privacy and at the same time provide useful data to the demanding agencies. In this paper, we presented an „optimized K-anonymization algorithm which is capable of handling a novel attack i.e. „mutual friends attack as well as „degree attacks while publishing social network data. The algorithm is based upon adding and deleting mutual friends of the nodes which are taken as noisy nodes. The proposed algorithm thus achieves the principles of K-anonymity while handling both the attacks. The experimental results show the utility of social networks after anonymization.

References

[1] Zhou., Bin., Jian Pei., WoShun Luk. (2008). A brief survey on anonymization techniques for privacy preserving publishing of social network data. *ACM Sigkdd Explorations Newsletter* 10 (2)12-22.

[2] Sun., Chongjing., S. Yu Philip., Xiangnan Kong., Yan Fu. (2013). Privacy preserving social network publication against mutual friend attacks. *In*: Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on, 883-890. IEEE,.

[3] Friends of Twitter user. @GURPREETSGREWAL .https://twitter.com/GURPREETSGREWAL/following. Last accessed on Jan 2017.

[4] Friends of Twitter user. @gurditS. https://twitter.com/gurditS/following. Last accessed on January 2017.

[5] Sweeney, Latanya. (2002). k-anonymity: A model for protecting privacy. (2002). *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10 (05) 557-570.

[6] Zhou, Bin., Jian Pei. (2008). Preserving privacy in social networks against neighborhood attacks. *In*: Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on, 506-515. IEEE.

[7] Wu, X., Ying, X., Liu, K.,Chen. L. (2009). A survey of algorithms for privacypreserving social network analysis. *Managing and Mining Graph Data. Kluwer Academic Publishers* (2009).

[8] Hay., Michael., Gerome Miklau., David Jensen., Philipp Weis., Siddharth Srivastava. (2007). Anonymizing social networks. 2

[9] Hay, Michael, Gerome Miklau, David Jensen, Don Towsley, and Philipp Weis. (2008) Resisting structural re-identification in anonymized social networks. *In*: Proceedings of the VLDB Endowment 1, (1) 102-114.

[10] Zou, Lei, Lei Chen, and M. Tamer Özsu. (2009). K-automorphism: A general framework for privacy preserving network publication. *In*: Proceedings of the VLDB Endowment 2, (1) 946-957.

[11] Liu., Kun., Evimaria Terzi. (2008). Towards identity anonymization on graphs. *In*: Proceedings of the 2008 ACM SIGMOD international conference on Management of data, p. 93-106. ACM.

[12] Ying., Xiaowei., Xintao Wu. (2008). Randomizing social networks: a spectrum preserving approach. *In*: Proceedings of the 2008 SIAM International Conference on Data Mining, 739-750. Society for Industrial and Applied Mathematics.

[13] Vuokko., Niko., Evimaria Terzi. (2010). Reconstructing randomized social networks. *In*: Proceedings of the 2010 SIAM International Conference on Data Mining, 49-59. Society for Industrial and Applied Mathematics.

[14] Wu, Leting, Xiaowei Ying, and Xintao Wu. Reconstruction from randomized graph via low rank approximation. *In*: Proceedings of the 2010 SIAM International Conference on Data Mining, 60-71. Society for Industrial and Applied Mathematics.

[15] Hay, Michael, Chao Li, Gerome Miklau, and David Jensen. Accurate estimation of the degree distribution of private networks. *In*: Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on, 169-178. IEEE.

[16] Zheleva, Elena., Lise Getoor. (2008). Preserving the privacy of sensitive relationships in graph data. *In*: Privacy, security, and trust in KDD, p. 153-171. Springer Berlin Heidelberg.

[17] Campan, A., TrutaTM. (2008). A Clustering Approach for Data and Structural Anonymity in Social Networks. 2nd ACM SIGKDD International Workshop on Privacy, Security, and Trust in KDD (PinKDD'08) (2008) r54.

[18] Machanavajjhala, Ashwin, Kifer, Daniel., Gehrke, Johannes., Venkitasubramaniam, Muthuramakrishnan. (2007). l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1 (1) 3.

[19] Prasad, Ajay, G. K., Panda, A., Mitra, Arjun Singh., Deepak Gour. (2010). Applying l-Diversity in anonymizing collaborative social network. *arXiv preprint arXiv:1007.0292*.

[20] Li, Na, Nan Zhang., Sajal K. Das. (2011). Relationship privacy preservation in publishing online social networks. *In*: Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on, 443-450. IEEE.

[21] Kavianpour., Sanaz., Zuraini Ismail., Amirhossein Mohtasebi. (2011). Preserving Identity of users in Social Network Sites By Integrating Anonymization and Diversification Algorithms. *International Journal of Digital Information and Wireless Communications (IJDIWC)* 1 (1) 32-40.

[22] Tripathy., Bala Krushna., Anirban Mitra. (2012). An algorithm to achieve k-anonymity and l-diversity anonymisation in social networks. *In*: Computational Aspects of Social Networks (CASoN), 2012 Fourth International Conference on, 126-131. IEEE.

[23] Yuan, Mingxuan, Lei Chen, S. Yu Philip, and Ting Yu. (2013). Protecting sensitive labels in social network data anonymization. *IEEE Transactions on Knowledge and Data Engineering* 25 (3) 633-647.

[24] Bhagat, Smriti, Cormode, Graham, Krishnamurthy, Balachander., Srivastava, Divesh .(2009) Class-based graph anonymization for social network data. *Proceedings of the VLDB Endowment* 2 (1) 766-777.

[25] Cormode, Graham., Divesh Srivastava., Ting Yu., Qing Zhang. (2008). Anonymizing bipartite graph data using safe groupings. *Proceedings of the VLDB Endowment* 1 (1) 833-844.

[26] Tassa., Tamir., Dror J. Cohen. (2013). Anonymization of centralized and distributed social networks by sequential clustering. *IEEE Transactions on Knowledge and Data Engineering* 25 (2) 311-324.

[27] NodeXL. http://nodexl.codeplex.com/. Last Accessed on April 2015.

[28] APL. https://networkx.github.io/documentation/latest/reference/generated networkx.algorithms.shortest_paths.generic .average_shortest_path_length.html. Last Accessed on January, 2016.

[29] De Maesschalck, Roy, Delphine Jouan-Rimbaud, and Désiré L. Massart. (2000). The mahalanobis distance. *Chemometrics and intelligent laboratory systems,* 50 (1) 1-18.

[30] Ford., Roy., Traian Marius Truta., Alina Campan. (2009). P-Sensitive K-Anonymity for Social Networks. *DMIN* 9 (2009): 403-409.