Anonymized Social Networks Community Preservation

Jyothi Vadisala
Dept. of Computer Science & Systems Engineering
College of Engineering(A), Andhra University
Visakhapatnam, India

Valli Kumari Vatsavayi
Dept. of Computer Science & Systems Engineering
College of Engineering(A), Andhra University
Visakhapatnam, India

***Abstract*—Social Networks have been widely used in the society. Most of the people are connected to one another, communicated with each other and share the information in different forms. The information gathered from different social networking sites is growing tremendously in large volumes of various research, marketing and other purposes which is creating security and privacy concerns. The gathered information contains some sensitive and private information about an individual, such as the relationship of an individual or group information. So, to protect the data from unauthorized users the data should be anonymized before publishing. In this paper, we study how the *k*-degree and *k*-NMF anonymized methods preserve the existing communities of the original social networks. We use an existing heuristic algorithm called Louvian method to identify the communities in social networks. We conduct the experiments on real data sets and compare the performances of the two anonymized social networks for preservation of communities of the original social networks.**

***Keywords*—Community; anonymity; degree; social network**

I. Introduction

Social networks are ubiquitous these days and are widely used for communication. The people are connected, whether near or far, anyone can be connected through social networks to anyone they want to and share the information like images, videos and text, etc. This data is published for various research purposes. Facebook, Twitter, Goggle are the best examples of social media where people share their information. These social networks must provide the privacy to their members and a privacy policy regarding how the collected data is used and published for various purposes. To protect the privacy of individuals the data must be anonymized before publishing the data. There are different anonymization algorithms which anonymizes the data. Most of the social network data are represented by graphs so there is no standard anonymization method which protects the privacy of individuals. In general, the privacy protection either identity of individuals, the relationship of individuals and the node content of their network. There are different anonymization methods and are applicable for appropriate privacy risks like anonymization via modification of the original graph, anonymization via clustering and differential privacy, etc. In this paper, we study how well the anonymized networks preserve the existing communities of the initial networks.

The communities of a social network mean groups of nodes which have similar characteristics or properties. There are different community detection algorithms presented in the paper [1]. In this paper, we use a heuristic algorithm called a Louvian method [2] based on modularity optimization. The modularity function has two values either positive or negative. The positive values indicate the presence of community structure possibilities. We follow a two steps to study how well the anonymized networks preserve the communities of the initial social networks. First, the initial network is anonymized by the two approaches, i.e *k*-degree anonymization and *k*-NMF anonymization. Second, we apply Louvian method to detect the communities from the anonymized networks and compare the two methods of preservation of communities of the initial networks by conducting experiments on real data sets.

II. Related Work

Several studies address the need of anonymizing the social networks to protect the privacy of individuals. Most of the prior work focus on preserving the structural properties between the original and anonymized social networks. The complete survey of existing social networks anonymization methods and the other privacy issues of the social networks is covered in [3]. An another important study of social networks is that of identifying communities in the network. Generally, communities are groups where we can identify the groups of interacting the nodes and the relationship between them like the friends group who studied in the same school or working in the same company, etc. There are so many papers which discussed how to detect the communities from the social networks. There are different algorithms are used to detect the communities. In Moradi and Olovsson et al. [4] used large e-mail networks to experimentally evaluate the qualitative performance of several community detection algorithms. In Malliaros and Vazirgiannis [5] suggested a methodology-based taxonomy to classify the different community detection approaches for directed graphs. The Ruan and Zhang [6] proposed a modularity measure to assess the quality of community structures. To compare the different community structures the modularity measure is well used. A larger modularity value means stronger community structures. The optimization of modularity measure is proposed in Newman [7], Duch and Arenas [8].

In this paper, we study the two graph modification approaches (*k*-degree and *k*-NMF anonymizations) and we focus on how these methods preserve communities of the original social network. To conduct experiments we consider the publicly available data sets and compare the results for the both

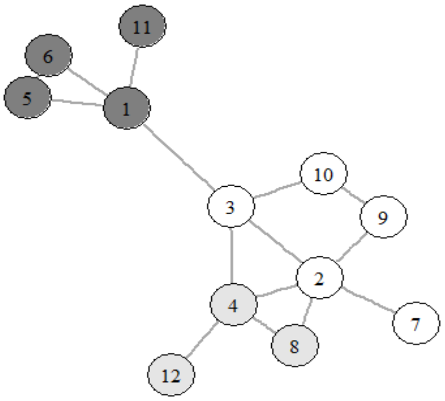


Fig. 1. Initial social network(G_1).

anonymized methods.

III. MODELS FOR SOCIAL NETWORK ANONYMITY

In this section we present the two anonymization techniques k-degree anonymity and k-NMF anonymity and we focus on the preservation of communities based on the structure of social networks. The process of anonymization is also based on the social network structural properties. Generally, a social networks are modelled as a graph $G = (V, E)$ where V is the set of vertices and E is the set of edges which represents the relationship between these vertices. In this paper, we consider the social network G as a simple undirected graph and the intruder knows the structure of the network and able to identify the individuals along the sensitive information due to the unique structure of the social network data. Fig. 1 shows the example of a social network which has 12 nodes and 15 edges.

A. k-degree anonymity

k-degree anonymity is the extension of well-known k-anonymity model where the intruder has the knowledge of the vertex degree to breach the identity of vertices. This method is a vertex based anonymization technique where there is at least $k - 1$ other vertices have the same degree. Liu, K.Terzi et al.[9] created an initial algorithm and proposed a k-degree anonymous network based on the degree property of the network. In this paper, we consider a Fast k-degree Anonymization Algorithm (FKDA) which is proposed by Lu et al.[10].

FKDA is a greedy algorithm in which the social network is anonymized by edge addition to the network until the network is k-degree anonymous. FKDA is a two step process, in Step 1 the vertices of original network is separated into several groups. Step 2, select each group and anonymize by adding edges to the vertices of the same group until all the vertices have the same degree in that group. If the group does not achieve the anonymization by edge creation, then it adds the edges by the relaxed edge addition method in which the vertices in that group are anonymized by connecting to other vertices in the graph rather than the same group. But the relax edge addition method may destroy the previously anonymized groups and the whole process will be restarted. The worst case time complexity of performing this approach is $\mathcal{O}(V^2)$ where V is the total number of vertices in the network. The Fig. 2 shows the anonymization of the graph $G^{'}{}_1$ using FKDA technique. The dashed edges represent the newly added edges by FKDA algorithm. The network has 3 nodes has degree 5, 2 nodes have degree 4 and 7 nodes has degree 3 so the network satisfies the 2-degree anonymity where $k = 2$. The details of the algorithm are specified in [10].

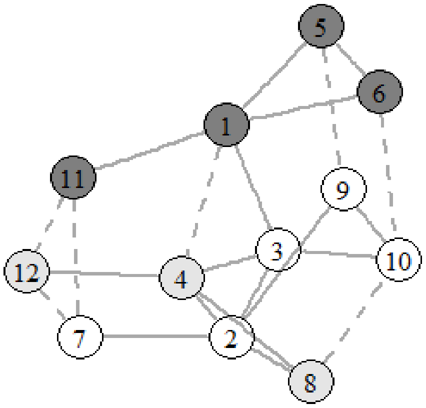


Fig. 2. k-Degree anonymized social network($G_2^{'}$).

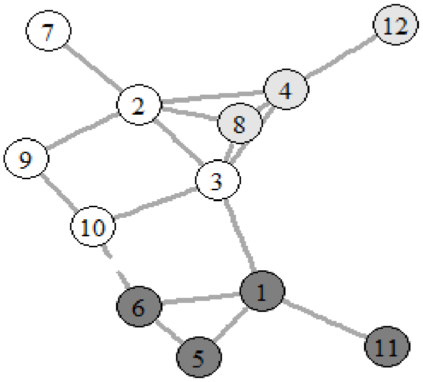


Fig. 3. k-NMF anonymized social network($G_3^{'}$).

B. k-NMF anonymity

In this method, we anonymize the original graph only by edge addition. The intruder has the background knowledge of the number of common vertices of an edge. This method is an edge based anonymization technique where there are at least k edges in a group has the same count of the number of common vertices. In k-NMF anonymization [11] first, we group the more than k edges and second, each edge of this group is anonymized by the breadth first search method. The edges are updated dynamically in the edge list because some new edges are added into the network. Therefore,when adding one edge will affect the count of the number of common friends of another edge or more edges. So it has to follow the anonymized triangle preservation principle which aims to preserve the already anonymized edges neither creating some additional anonymized triangles by edge addition or destroy by the edge deletion. This preservation leads to avoid repeatedly anonymizing the same edges. The Fig. 3 shows an example of *k*-NMF anonymization process for the graph G_1.

The *k*-NMF anonymization problem can be seen as a parallel of *k*-degree anonymization problem. As in *k*-degree anonymization process needs more number of edge additions than *k*-NMF anonymity so most of the structural properties of a graph will be preserved by the *k*-NMF anonymization algorithm. This significant difference in the privacy protection of individuals between the two methods leads us to the *k*-NMF anonymity will preserve the communities of the original social networks than the *k*-degree anonymity model. This can be explained by conducting experiments on real data sets.

IV. Community Detection

In this paper, we focus on preserving communities by anonymized social networks. Generally, identifying a community in complex networks is a universal problem and has consequently been raised in many domains, leading to different solutions. Most of the community detection methods rely on Newmans modularity to assess the quality of their results. The community detection algorithms are grouped as hierarchical, optimization and others. In hierarchical approach the result is a tree of the communities which is represented as a dendrogram. The hierarchical method consists of two approaches, i.e Agglomerative and Divisive. The optimization-based approaches use a Newmans modularity measure to calculate the quality of a network partition. The algorithm consists of two steps. In Step 1, processing several partitions of the network either randomly or by a fitting function and in Step 2 based on quality measure choose the best nodes and this algorithm is modified to get the better quality. Most of the optimization algorithms have used a modularity measure because it is a costly measure to process [12], [13], [14]. Other algorithms use a clustering principle [15] [16] and also find the overlapping communities, i.e. one node may be a part of several communities at once [17]. In Derneyi et al. [18] used the latent space approach to process the probability for a node to belong to a community. In this paper, we use a Louvain community detection method for detecting the communities from the original and anonymized social networks. In this method each node is assigned to one community. Then the modularity gain of each community is maximized by moving nodes between those communities. This step is stopped when there is no change in modularity gain with the movement of nodes. After this process the network obtained from the first step is used and a weighted network is created. In this weighted network, one node represents a community from the original network, and weights are added to edges to represent the number of original edges that are collapsed into a super edge. After the completion of this step again the first step is implemented. This process repeated iteratively until the modularity gain is maximized. Communities obtained by the Louvain method for a graph G_1 are shown in Fig. 1. The color of a vertex represents the community they belong.

TABLE I. Preservation at Community Level (*k*-Degree Anonymity)

Community	Communities in G_1	Communities in $G_2^{'}$	Preservation of Community(%)
1	{1,5,6,11}	{1,5,6,9,10}	33%
2	{4,8,12}	{2,3,4,8}	66%
3	{2,3,7,9,10}	{7,11,12}	20%

TABLE II. Preservation at Community Level (*k*-NMF Anonymity)

Community	Communities in G_1	Communities in $G_2^{'}$	Preservation of Community(%)
1	{1,5,6,11}	{1,5,6,11}	100%
2	{4,8,12}	{2,3,4,7,8,12}	100%
3	{2,3,7,9,10}	{9,10}	40%

A. *Community Preservation*

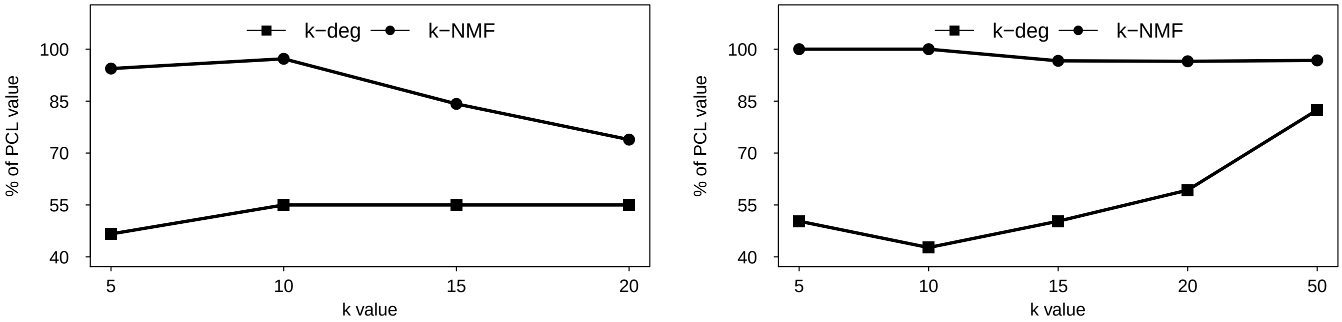
In this section, we estimate the preservation of communities by the anonymizaed social networks and compare with the communities of the original social networks. We compute the communities of the anonymized social networks and original networks using Louvian method and compare the results between the anonymized and original networks using two different approaches.

1) Preservation at Community Level (PCL)*:* In this, we count how many vertices have remained in the same community after the anonymization process. The preservation of communities by two anonymization methods is shown in Table I and Table II. The percentage of preservation of communities is calculated for each community of the graph G_1 with the corresponding community of Graphs $G_2^{'}$ or $G_3^{'}$ that contain the maximum number of vertices from the original community. The *PCL* value for a network will be calculated by the average of the results for the percentage of preservation of each community.

The percentage of preservation of each community in the initial social network and anonymized social networks is shown in Tables I and II. For example the percentage preservation for the second community from Table I i.e. {4,8,12}, the best match is the community {2,3,4,8} and the percentage of preservation is $\frac{2}{3}$. To measure the percentage of preservation for the network is the sum of all results of the percentage of preservation of communities divided by the total number of communities. The *PCL* values for the two anonymized social networks is given below:

- $PCL(G_1, G_2^{'}) = 39.66\%$
- $PCL(G_1, G_3^{'}) = 80\%$

2) Preservation of Community at Node Level (PCNL): In this section, we estimate the preservation of communities at each node individually. We compare the community of each node at original network and anonymized network. Consider the initial social network as $G = (V, E)$ and the anonymized social networks as $G' = (V, E')$. The set of nodes $V = v_1, v_2, \ldots v_n$ and the $Com(v_i)$ and $Com'(v_i)$ represent the node community at original and anonymized social networks respectively. The community preservation for each node v_i is



(a) karate social network

(b) powergrid social network

Fig. 4. Percentage of preservation of community level ($\%PCL$).

TABLE III. Preservation of Community at Node Level (k-Degree Anonymity)

Node	$Com(v_i)$	$Com'(v_i)$	$PCNL(v_i)$
1	{1,5,6,11}	{1,5,6,9,10}	50%
2	{2,3,7,9,10}	{2,3,4,8}	28.57%
3	{2,3,7,9,10}	{2,3,4,8}	28.57%
4	{4,8,12}	{2,3,4,8}	40%
5	{1,5,6,11}	{1,5,6,9,10}	50%
6	{1,5,6,11}	{1,5,6,9,10}	50%
7	{2,3,7,9,10}	{7,11,12}	14.28%
8	{4,8,12}	{2,3,4,8}	40%
9	{2,3,7,9,10}	{1,5,6,9,10}	25%
10	{2,3,7,9,10}	{1,5,6,9,10}	25%
11	{1,5,6,11}	{7,11,12}	16.66%
12	{4,8,12}	{7,11,12}	20%

TABLE IV. Preservation of Community at Node Level (k-NMF Anonymity)

Node	$Com(v_i)$	$Com'(v_i)$	$PCNL(v_i)$
1	{1,5,6,11}	{1,5,6,11}	100%
2	{2,3,7,9,10}	{2,3,4,7,8,12}	37.5%
3	{2,3,7,9,10}	{2,3,4,7,8,12}	37.5%
4	{4,8,12}	{2,3,4,7,8,12}	50%
5	{1,5,6,11}	{1,5,6,11}	100%
6	{1,5,6,11}	{1,5,6,11}	100%
7	{2,3,7,9,10}	{2,3,4,7,8,12}	37.5%
8	{4,8,12}	{2,3,4,7,8,12}	50%
9	{2,3,7,9,10}	{9,10}	40%
10	{2,3,7,9,10}	{9,10}	40%
11	{1,5,6,11}	{1,5,6,11}	100%
12	{4,8,12}	{2,3,4,7,8,12}	50%

the number of nodes common in both $Com(v_i)$ and $Com'(v_i)$ divided by the at least one of these two communities.

$$PCNL(v_i) = \frac{|Com(v_i) \cap Com'(v_i)|}{|Com(v_i) \cup Com'(v_i)|} \tag{1}$$

Where $|V|$ represents the number of elements in set V. The final *PCNL* value is calculated as the sum of all individual preservation of community node values divided by the total number of nodes in the network is shown below:

$$PCNL(G, G') = \frac{\sum_{i=1}^{n} PCNL(v_i)}{n} \tag{2}$$

The preservation of community at node level for the two anonymized networks of Fig. 2 & 3. is shown in Tables III & IV. To illustrate this computation, let us consider the node 4 from Table III. The initial community for the node 4 is {4,8,12} and the k-degree anonymized community is {2,3,4,8}. By observation, there are two nodes common in these two sets i.e., {4,8} and 5 nodes in their union of sets i.e {2,3,4,8,12}. So the *PCNL* value for node 5 is $\frac{2}{5}$. The final preservation of community at node level for each anonymized social network with respect to original network is shown below:

- $PCNL(G_1, G_2^{'}) = 32.34\%$
- $PCNL(G_1, G_3^{'}) = 61.87\%$

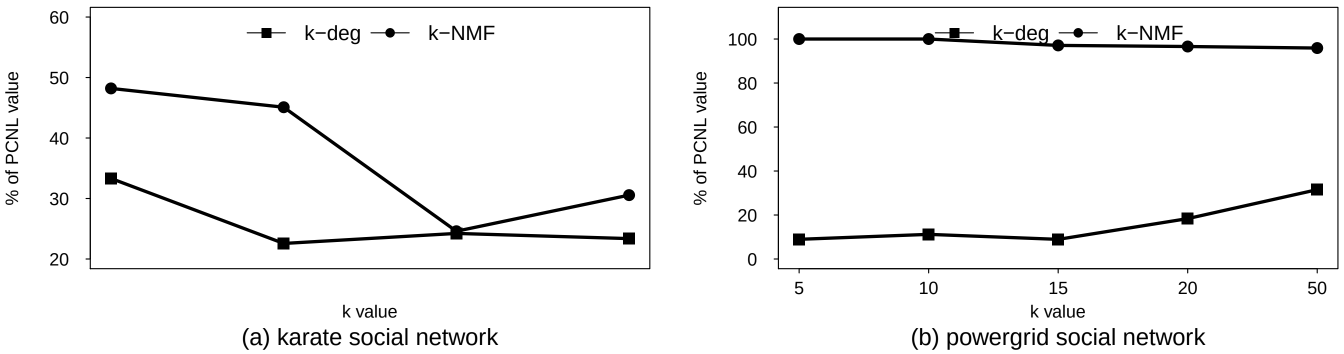
V. Experimental Results

In this section, the following publicly available data sets are used for the preservation of communities between original and anonymized social networks.

- Zacharys karate club is a small undirected friendship relation social network. It has 34 nodes, 78 edges and 4 communities.
- A Power grid is an undirected, unweighted network representing the topology of the western states power grid of the united states. This network consists of 4,941 nodes, 6,594 edges and 40 communities.

We performed different steps to measure the preservation of communities. In step1, first we consider the above initial networks, and anonymize these networks by the two anonymization methods (FKDA,k-NMF) using several anonymity values of k i.e 5, 10, 15, 20 and 50. Next we calculated the communities of the initial networks, and each anonymized network using a Louvian method in R programming. Finally, we compute the preservation of communities using *PCL* and *PCNL* approaches and plot the average results of *PCL* and *PCNL* values for the above networks.

Fig. 4 represents the percentage of preservation of communities of karate and power grid social networks. In both the networks, a k-NMF method preserves the communities



(a) karate social network

(b) powergrid social network

Fig. 5. Percentage of preservation of community at node level ($\%PCNL$).

of original network very well. In k-NMF anonymization preserved the communities of the original network well when k is small but as the k value increases the preservation of communities is decreased for the karate social network dataset. Similarly, Fig. 5 represents the percentage of preservation of communities at node level is also well preserved by k-NMF anonymization method for both the data sets. But at the $k = 15$ the *PCNL* value decreased for karate network. Based on the above results, we conclude that the k-NMF algorithm preserves the communities of original network very well using both *PCL* and *PCNL* measures.

VI. Conclusion

In this work, we focused on how well the anonymized social networks preserve the communities of the original social networks. We analyzed k-degree anonymization model where the adversary identifies a vertex based on the degree of a vertex as a background knowledge, whereas a k-NMF model, the adversary identifies an edge, based on the number of common friends of the connected edge as a background knowledge. Our results show that the k-NMF model preserves the very well communities than the k-degree model. However, there are several future directions have to be considered. First, while anonymizing the social networks if the number of vertices are increased, then how well the anonymized networks will preserve the communities of the original network. Second, while anonymizing if the number of communities are increases or decreases how the communities are preserved in large complex networks. Our method does not discuss these situations, therefore we plan in future to create a more robust way of comparing community preservation.

References

[1] Gunce Orman, Vincent Labatut. A Comparison of Community Detection Algorithms on Artifi- cial Networks. Discovery Science (DS), 2009, Porto, Portugal. Springer, 5808, pp. 242-256, 2009.

[2] Blondel, Vincent D., et al. "Fast unfolding of communities in large networks." Journal of statistical mechanics: theory and experiment 2008.10 (2008): P10008.

[3] Jyothi Vadisala and Valli Kumari Vatsavayi, Challenges in Social Network Data Privacy, International Journal of Computational Intelligence Research (IJCIR), vol -13, pp. 965-979, 2017.

[4] F Moradi, T Olovsson, P Tsigas, An of community detection algorithms on large-scale email traffic, in: SEA. Berlin/Heidelberg: Springer; 2012; 283294.

[5] F.D Malliaros, M Vazirgiannis, Clustering and community detection in directed networks: a survey, CoRR, abs/1308.0971, 2013.

[6] J. Ruan and W. Zhang, An Efficient Spectral Algorithm for Network Community Discovery and Its Applications to Biological and Social Networks, Proc. Seventh IEEE Intl Conf. Data Mining (ICDM 07), pp. 643-648, Jan. 2007.

[7] M. Newman, The Structure and Function of Complex Networks, SIAM Rev., vol. 45, no. 2, pp. 167-256, 2003.

[8] Jordi Duch, Alex Arenas, Community detection in complex networks using extremal optimization, Phys, Rev E72, 027104, August 2005.

[9] Liu, Kun, and Evimaria Terzi. "Towards identity anonymization on graphs." Proceedings of the 2008 ACM SIGMOD international conference on Management of data. ACM, 2008.

[10] Lu, Xuesong, Yi Song, and Stphane Bressan. "Fast identity anonymization on graphs.", proceedings of the 23rd International Conference on Database and Expert Systems Applications. Springer Berlin/Heidelberg, pp. 281-295, 2012.

[11] C. Sun, P. S. Yu, X. Kong and Y. Fu, "Privacy Preserving Social Network Publication against Mutual Friend Attacks," 2013 IEEE 13th International Conference on Data Mining Workshops, Dallas, TX, 2013, pp. 883-890.

[12] Duch, J., Arenas, A.: Community detection in complex networks using extremal optimization. Physical Review E 72 (2005) 027104 (3).

[13] Reichardt, J., Bornholdt, S.: Detecting Fuzzy Community Structures in Complex Networks with a Potts Model. Physical Review Letters 93 (2004) 218701.

[14] Clauset, A., Newman, M.E.J., Moore, C.: Finding community structure in very large networks. Physical Review E 70 (2004) 066111.

[15] Falkowski, T., Barth, A., Spiliopoulou, M.: DENGRAPH: A Density-based Community Detection Algorithm. IEEE/WIC/ACM International Conference on Web Intelligence (2007) 112-115.

[16] Liu, Y., Wang, Q., Wang, Q., Yao, Q., Liu, Y.: Email Community Detection Using Artificial Ant Colony Clustering. Advances in Web and Network Technologies, and Information Management. Springer, Berlin / Heidelberg (2007) 287-298.

[17] . Hoff, P., Raftery, A., Handcock, M.: Latent space approaches to social network analysis. Journal of the American Statistical Association 97 (2002) 1090-1098.

[18] Derenyi, I., Palla, G., Vicsek, T.: Clique percolation in random networks. Physical Review Letters 94 (2005).