Penerbit Akademia Baru

Journal of Advanced Research in Computing and Applications

Journal homepage: www.akademiabaru.com/arca.html

ISSN: 2462-1927

Anonymizing published social network data

Open Access

Emad Elabd[1,*], Hatem AbdulKader[1], Waleed Ead[2]

[1] Faculty of computers and information, Menoufia University, Gamal Abd El-Nasir, Qism Shebeen El-Kom, Shebeen El-Kom, Menofia Governorate, Egypt

[2] Faculty of Computers and Information, Beni-Suef University, Egypt

ARTICLE INFO

Article history:
Received 5 June 2017
Received in revised form 10 July 2017
Accepted 15 July 2017
Available online 4 December 2017

Keywords:
Data publishing, privacy preserving, online social networks, background knowledge, anonymization, frequent pattern mining

ABSTRACT

Interpersonal organization information give significant data to organizations to better comprehend the attributes of their potential clients as for their groups. Yet, offering informal community information in its crude structure raises serious security concerns. An adversary may attack the privacy of certain victims easily by collecting local background knowledge about individuals in a social network such as information about its neighbours. Subsequently, many anonymization algorithms were proposed to solve such issues. In this paper, a secure k-anonymity algorithm to protect published data against such named structural attacks (e.g. Degree Attack and subgraph attack) is proposed. Experimental results showed that the anonymized Online Social Networks (OSNs) can preserve much of the characteristics of original OSNs as a tradeoff between privacy and utility.

1. Introduction

Last years have seen an exponential growth in the Social Network Systems (SNSs) such as Facebook and Twitter, with different concerning stores of the privacy and security of appearing repeatedly in mainstream media. According to Boyd and Ellison [1], a "social network site" is characterized by three functions. Firstly, the social network site allows users to create a public or semi-public use profile of themselves. Secondly, it provides formal facilities for their users to express their relationships with other social users such as the friend lists that may typically reflect their existing social connections. Lastly, social users may traverse their relationships in order to explore the space of their user profiles.

Moreover, online social networks (OSNs) such as Facebook, Twitter, and Myspace provide information about individuals in some population and show the links between them. Such links may describe the relations of collaboration, friendship, and correspondence. Some real OSNs are

[*] *Corresponding author.*
E-mail address: emadqap@gmail.com (Emad Elabd)

complex and contain a huge set of information. These social networks become an important data source that can be published for different analysis purposes [2].

Many of real-world OSNs contains sensitive information and serious privacy [3-6]. As a result, research on preserving the privacy of published OSNs data has begun to receive more attention [7].

The goal of social network analysis is to uncover hidden social patterns. In social network analysis, the relationships between social individuals in the social networks are regarded more important and beneficial than the attributes information of social individuals [8]. As a result preserving privacy in publishing OSNs data becomes an important concern.

1.1 Adversary Background Knowledge

An attacker usually relies on his/her background knowledge to re-identify individual's vertices and learn the link relationships between individuals from the released anonymized graph. These assumptions of the adversary's background knowledge play a critical role in modelling privacy attacks and developing methods to protect privacy in social network data. The attacker's background knowledge [8] can be: attributes of vertices, specific link relationships between some target individuals, vertex degrees, neighborhoods of some target individuals, embedded subgraphs, and graph metrics (e.g., centrality, closeness, betweenness). For example, Liu and Terzi [9] considered vertex degrees as background knowledge of the adversaries to breach the privacy of target individuals. The work proposed in [10] used neighborhood structural information of some target individuals. While authors in [8, 11] proposed the use of embedded subgraphs and Ying *et al*. [12] exploited the topological similarity/distance to breach the link privacy.

The adversary may use his background knowledge to attack the privacy of some victims by collecting some local knowledge about the target individual vertices in a social network. To clear, consider a synthesized social network of friends as shown in Figure 1. Each vertex in the network represents an individual. Two individuals are friends if they are linked by an edge.

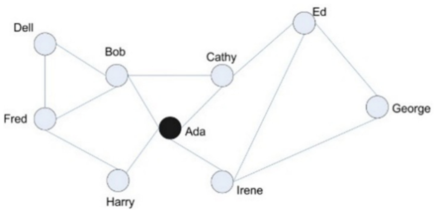

Fig. 1. Synthesized Social Network of friends

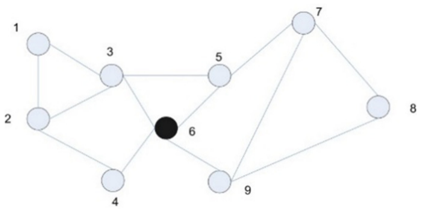

Fig. 2. A naïve anonymized Social Network

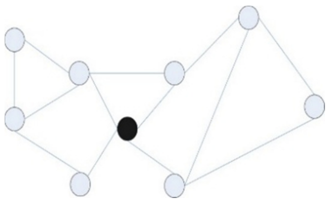

Fig. 3. A naïve anonymized Social Network without node identification

The SN can be naïvely anonymized by assigning a random ID to each node, e.g. Dell is assigned to 1, Bob is assigned to 2 and so on, as shown in figure 2. Another way to naïve anonymization the SN by removing the node identification as shown in figure 3. Suppose that this social network is to be published. To preserve the privacy, it is not sufficient to assign a random ID to each identity; figure 2; or remove all identities of vertices; figure 3. If an adversary, unfortunately, has some knowledge about the subgraph of locating an individual, the privacy may still be leaked [10].The adversary may use his/her background knowledge such as identity subgraph to attack the privacy of some victims.

1.2 Research Problem and Contribution

Many anonymization approaches were proposed to solve privacy preserving of published social networks [3, 10, 13-17]. Most of these approaches do not guarantee the privacy preserving of identities from attacker's frequent pattern (subgraph) attack as background knowledge. Social graph frequent patterns are sub-graphs that may found in a collection of social graphs or a single massive social graph with a frequency no less than a predefined user threshold [18].
The contribution of this paper may be summarized as follow:

- Developing an anonymization technique to preserve the privacy of released social network individuals.
- Providing an effective privacy preserving algorithm with a reasonable tradeoff between privacy and the information utility of the original social graph.
- Performance evaluation measures such as shortest path length, cluster coefficient, and degree distribution [10] will be conducted on real life datasets

The rest of paper will be organized as follow: Section 2 surveyed most related work in preserving the privacy of released SNs. Section 3 contains the proposed secure k-anonymity algorithm followed by experimental results and conclusion in Section 4 and 5 respectively.

2. Related Work

Li *et al*. [19] proposed k-anonymity technique to preserve privacy attack based on frequent patterns; one of the most important kinds of knowledge required for marketing and consumer behavior analysis [18]. In addition, Anusha *et al*. [15] present a framework that provides privacy to individuals in a social network against the adversary from frequent patterns. Their anonymization algorithm is based on the Degree Smoothing method. Such approach suffers from two limitations. The first limitation of this approach is that it only deals with 1-neighborhood such that if an attacker has the background knowledge about 1- neighborhood. The k-anonymity social network still suffers from neighborhood attacks. The second limitation, if the attacker has both the structural background knowledge of the social network and a partial label information of the target individual, such approach is insufficient under such attack. In addition, it doesn't provide information utility measures such as cluster coefficient or shortest path length between social actors. An adversary can reveal the users’ trusted social contacts. Mitta *et al.* [13] and Gartner, *et al.* [20] have proposed anonymization techniques to preserve the trusted user's contacts. The algorithm was based on the use of the degree constrained subgraph satisfaction problem on the complement of the input graph. In addition, it does not preserve the social frequent pattern analysis after anonymizing the social links. Ninggal *et al.* [16] present a type of vertex re-identification attack model called neighborhood-pair attack. This attack utilizes the information about the local communities of two connected vertices to identify the target individual. This work cannot protect the relationship (sensitive edge) attack. Zakrzadeh *et al.* [21] prevent the attribute disclosure attack without manipulating the graph structure. In this approach, the pattern of a specified user can easily reveal by an adversary. Karas *et al.* [22] and Prashanth *et al.*[23] published their graph in a form such that an adversary who possesses information about a node's neighborhood cannot safely infer its identity and its sensitive labels. The approaches in [21-24] suffer from preserving a subgraph pattern of an identity from the adversary frequent pattern knowledge. Campan *et al.* [25] proposed k-anonymity to preserve the social network community (subgraph) of a specified node not the entire social network nodes.

The previous related works proposed anonymization techniques but it does not guarantee protection to preserve the identity frequent pattern from the attacker's background knowledge. In the proposed algorithm we provide an effective privacy preserving technique to preserve the privacy of individuals under the sub-graph (pattern) attack. Moreover, it provides an effective way to preserve the different network properties such as network's degrees, cluster coefficients, and shortest path lengths.

3. K-Anonymity Algorithm

In this section, we demonstrate the secure k-anonymity algorithm solution for the anonymized social network (SN) graph.

Definition 1: A Social Network (SN) can be represented by a simple graph, G (V, E), where V is a set of vertices and E V x V, is a set of edges. A label function l maps a vertex or an edge to a label. V (G) or E (G) describes the vertex and edge set of G respectively. A social graph G’ = (V’, E’) is a sub-graph of G (V, E), denoted by G’ such that V’ , (u,v) ,and (u,v) .

The main idea for solution as follows: Given a SN graph G= (V, E), derive a released graph Gk= (Vk,Ek), Gk is secured K-anonymity, in which Gk={g1,g2,g3,…gk} with pairwise isomorphic gi and gj, .

Definition 2: A Secure K-anonymity SN: Let G= (V, E) be a given with unique node information, for each vertex (node). Let Gk to be the anonymized graph of G and Gk is secure anonymized with respect to G if for two individual targets vertices A and B with corresponding to the sub-graph attack that known by the attacker. For a given an anonymized Gk the attacker cannot determine which a pattern belongs to target victim A. In addition, given two sub-graph, the attacker cannot determine which pattern belongs to a target victim A or B with a probability not greater than 1/K.

The problem of privacy preserving in publishing a graph by the secure k-anonymity is defined as:

Definition 3: Given an original SN graph G= (V, E) and positive integer k, such that release an anonymized graph Gk= (Vk,Ek) to be published such that Vk=V; no dummy vertex node added. Gk is secure k-anonymity with respect to G. The anonymization cost from G to Gk is should be minimized.
Definition 4: Subgraph (graph pattern) Attack: an attacker knows the structure of the user and find the structure in the graph or find out which sub-graph match with the user structure in the network. In other words, the adversary may know a connected sub-graph Ga, and a vertex V in Ga that may belong to an individual A.

The secure k- anonymity published graph is based on the secure of anonymized subgraphs, as there are K different vertices a1, a2,...ak that may be mapped to A and k different vertices b1,b2,..bk that can be mapped to B, where and for.

The following example can demonstrate the above explanation, consider the following figure 4:

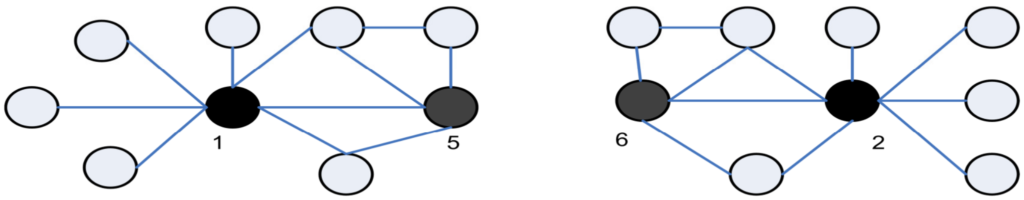

Fig. 4-a. Anonymized Graph G_k

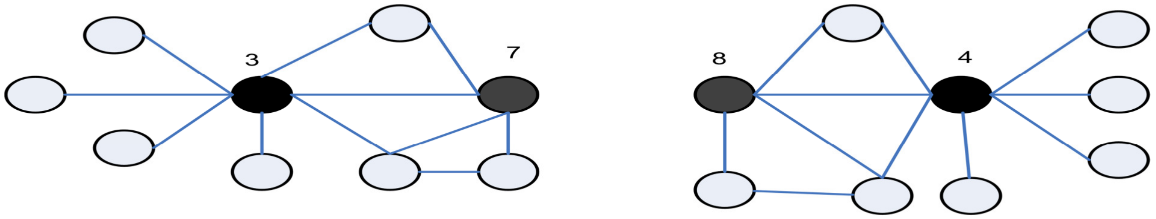

Fig.4- b. Anonymized Graph G_k

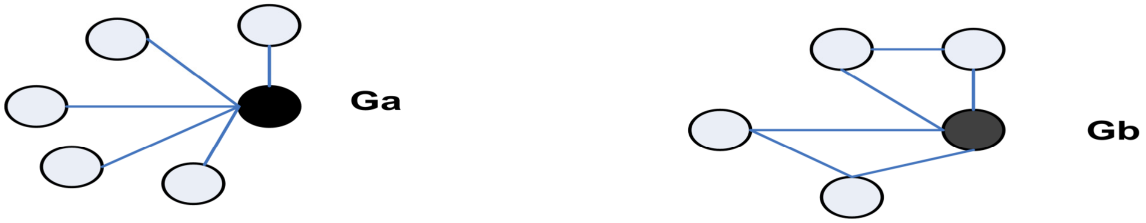

Fig.4-c. Sub-graph of A and B

Fig. 4. published *k*-anonymity graph

Figure 4, shows as an example of published secure k-anonymity graph. In figure 4c, the adversary attack with the two subgraph attacks Ga and Gb for the target individuals A and B respectively, which are two nodes in the graph. There are four vertices {1, 2} in Fig. 4a and {3, 4} in Fig. 4b that are linkable to A. While {5, 6} in Fig. 4a and, {7, 8} in Fig. 4b are linkable to B. The attacker can determine only that A and B are linked by an edge with a probability ¼. The proposed k-anonymity algorithm will lead to better privacy preserving of published SN individual’s data.

```
Algorithm 1: K-anonymity algorithm
Input: Graph G, K.
Output: an anonymized released graph $G_k$= {$g_1$, $g_2$,…$g_k$} of G.
   1. Traverse each vertex in G.
         a. Enumerate all connected K subgraph.
         b. If vertex set of these connected subgraphs can’t cover all vertices in G
               i. Enumerate such subgraphs.
                  % this means, its isolated components in G%
   2. Extract the vertices of K embedding to be transformed from G to the $g_i$’s in Gk.
      % No necessary to edge modification for anonymization with respect to such
      embedding %
   3. Remove this embedding from G.
   4. Anonymize $g_i$
         a. Enumerate g’s size that is isomorphic subgraphs.
         b. For each enumerated $g' \subset g$ .
               i. Find all embedding of g’.
              ii. Such that locate the embedding of these subgraphs within each
                  embedding.
                  % Rather than search the big graph G%
             iii. Let T to be a temporary table
                     1. Keep in T every subgraph g’ that have been processed a long
                        with embedding embd(g’) that have been uncovered.
                        % T will help us to check if g’ is used or not %
         c. Add edges in each $g_i$ for isomorphic pairwise subgraphs.
   5. Exit.
```

The proposed k-anonymity algorithm aims to generate and publish an SN graph Gk that consists of identity subgraphs. The set of nodes in G graph are partitioned into k-subgraph with the same number of vertices.

Definition 5: Given a pattern p = (V (p), E(p)), a simple overlap of occurrences g and g' of pattern p exists if g(E(p))∩g'(E(p)) ≠ ∅. The support of p is defined as the size of the maximum independent set (MIS) of the overlap-graph. A later study [18] proved that the MIS-support is anti-monotone.

Definition 6: Anti-monotonicity : a k subgraph is a frequent only if all of its sub-graphs are frequent[18] . For each k partition we ensure the each subgraph gi has a symmetry like the others gi's in other k partition by an edge addition or deletion to ensure the Anti-monotonicity of the k-subgraph.

Definition 7: A Frequent sub-graph: Given a set of graph datasets D = {G1, G2, . . . ,Gn} and a sub-graph g, the support of a sub-graph set of g is Dg = {Gi|g ⊆ Gi,Gi ∈ D}. The support of g = |Dg|/|D|. A frequent sub-graph is a sub-graph whose support is not less than a support threshold.

The frequent subgraph has shown to be a sounded in [10, 14]. In this work, the frequent subgraph has a high possibility to generate a set of connected subgraphs that will minimize the edge modifications needed for the graphs to be k-anonymity.

For better performance, we use the number of edges of a subgraph as a threshold to determine the number of subgraphs to be discovered. One way to determine that threshold is the average degree of G. A justification about determining such threshold is that many nodes in G may have this d degree and each form a potential anonymization subgraph with their d 1-neighbors.

To clear and summarize, we have a non-anonymized G that consists of multiple components (subgraphs), and a MAX threshold of number of anonymized subgraphs = k and the value of K partitions of the social graph. We enumerate the set of subgraphs with k edges. In addition, these subgraphs don’t cover the entire graph. Subsequently, do the following:

1. Determine the MIS of such subgraphs.
2. Determine the highest degree in each MIS.
3. Enumerate the number of graphs in each MIS with such degree.
4. Such graph and its MIS are entered into a temporary table T.

4. Experimental Results

Different real-life datasets are used for the proposed algorithm experimental analysis. We have extracted two datasets (dataset1 and dataset2) from large datasets such as EU email dataset with a different number of vertices 5000, 10000 respectively to test the proposed approach. We use different k values in the proposed algorithm to generate the anonymized graph for the different datasets. For short, we will demonstrate experimental results for k=10.

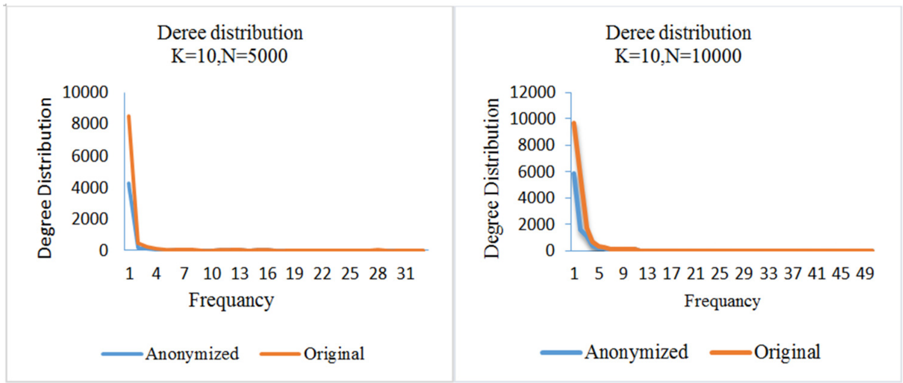

Fig. 5. Degree distribution (K=10)

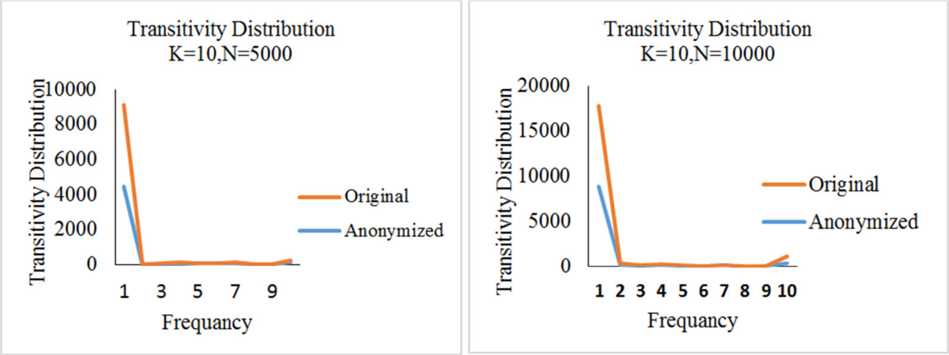

Fig. 6. Transitivity distribution (k=10)

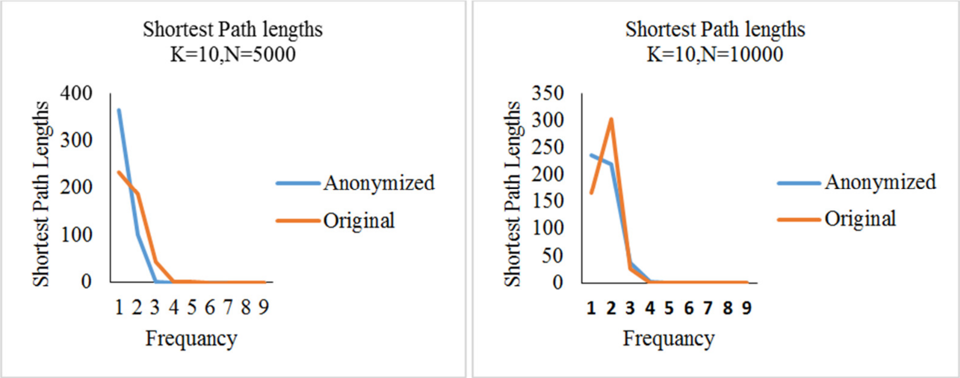

Fig. 7. Shortest path length distribution (k=10)

The following figure 5, shows the degree distribution for the original, anonymized (published) social graph. The figures show how the proposed algorithm keep and preserve the essential graph information such as degree distribution for different datasets.

As previously shown in the experimental results that the released secure k-anonymity graph will preserve the most essential properties such as degree distribution, transitivity distribution and shortest path distribution. As a result, released secure k-anonymity graph can answer aggregate quires with high accuracy. We have a tradeoff between the privacy preserving and social graph utility; the expected use of the social graph. The experiments analysis have been showed that high values of k in the anonymization of the original graph will lead to loss the expected social graph utility , so we must such the tradeoff between privacy preserving and the graph utility.

5. Conclusion

Publishing social network data is an important issue as these data may contain a treasure of information need to be sanitized. In this paper, we introduced a k- anonymity technique for preventing OSNs identities from adversary's background subgraph (pattern) attack. The proposed algorithm has preserved the essential graph utility information such as degree, transitivity and shortest path distributions for the original social graph against the anonymized published social graph. Building privacy preserving to other attacker's background knowledge will be interested. As an open issue will be taken in the information associated with each identity and labeled edges and show how such labeled information released securely can. Another issue is to propose a privacy preserving framework for publishing the social graph profiles data for OSN's users.

References

[1] Fong, Philip, Mohd Anwar, and Zhen Zhao. "A privacy preservation model for facebook-style social network systems." *Computer Security–ESORICS 2009* (2009): 303-320.

[2] Tassa, Tamir, and Dror J. Cohen. "Anonymization of centralized and distributed social networks by sequential clustering." *IEEE Transactions on Knowledge and Data Engineering* 25, no. 2 (2013): 311-324.

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

[4] Leskovec, Jure, Lars Backstrom, Ravi Kumar, and Andrew Tomkins. "Microscopic evolution of social networks." In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 462-470. ACM, 2008.

[5] Hay, Michael, Gerome Miklau, David Jensen, Don Towsley, and Philipp Weis. "Resisting structural re-identification in anonymized social networks." *Proceedings of the VLDB Endowment* 1, no. 1 (2008): 102-114.

[6] Backstrom, Lars, Cynthia Dwork, and Jon Kleinberg. "Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography." In *Proceedings of the 16th international conference on World Wide Web*, pp. 181-190. ACM, 2007.

[7] Adamic, Lada A., and Natalie Glance. "The political blogosphere and the 2004 US election: divided they blog." In *Proceedings of the 3rd international workshop on Link discovery*, pp. 36-43. ACM, 2005.

[8] Zhou, Bin, Jian Pei, and WoShun Luk. "A brief survey on anonymization techniques for privacy preserving publishing of social network data." *ACM Sigkdd Explorations Newsletter* 10, no. 2 (2008): 12-22.

[9] Liu, Kun, and Evimaria Terzi. "Towards identity anonymization on graphs." In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pp. 93-106. ACM, 2008.

[10] Zhou, Bin, and Jian Pei. "Preserving privacy in social networks against neighborhood attacks." In *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*, pp. 506-515. IEEE, 2008.

[11] Wu, Xintao, Xiaowei Ying, Kun Liu, and Lei Chen. "A survey of privacy-preservation of graphs and social networks." *Managing and mining graph data* (2010): 421-453.

[12] Ying, Xiaowei, and Xintao Wu. "On link privacy in randomizing social networks." *Knowledge and information systems* 28, no. 3 (2011): 645-663.

[13] Liu, Changhchang, and Prateek Mittal. "Linkmirage: How to anonymize links in dynamic social systems." *arXiv preprint arXiv:1501.01361* (2015).

[14] Deshpande, Amolika N. Patil Dr SP. "A Review on Privacy Priserving Data Publishing of Social Network." *network* 7, no. 8: 10.

[15] Anusha, K., and K. Venkata Ramana. "Degree Smoothing On Social Networks against Frequent Shared Patterns." *Int. J. Adv. Res. Sci. Technol. Volume* 4, no. 6 (2015): 435-439.

[16] Abawajy, Jemal, Mohd Izuan Hafez Ninggal, and Tutut Herawan. "Vertex re-identification attack using neighbourhood-pair properties." *Concurrency and Computation: Practice and Experience* 28, no. 10 (2016): 2906-2919.

[17] Kossinets, Gueorgi, Jon Kleinberg, and Duncan Watts. "The structure of information pathways in a social communication network." In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 435-443. ACM, 2008.

[18] C. C. Aggarwal, "Mining Graph Data," in Data Mining, 2015, pp. 557-587.

[19] Fung, Benjamin, Yan'an Jin, and Jiaming Li. "Preserving privacy and frequent sharing patterns for social network data publishing." In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 479-485. ACM, 2013.

[20] Chester, Sean, Jared Gaertner, Ulrike Stege, and S. Venkatesh. "Anonymizing subsets of social networks with degree constrained subgraphs." In *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*, pp. 418-422. IEEE, 2012.

[21] Zakerzadeh, Hessam, Charu C. Aggarwal, and Ken Barker. "Big Graph Privacy." In *EDBT/ICDT Workshops*, pp. 255-262. 2015.

[22] Song, Yi, Panagiotis Karras, Qian Xiao, and Stéphane Bressan. "Sensitive Label Privacy Protection on Social Network Data." In *SSDBM*, pp. 562-571. 2012.

[23] Song, Yi, Panagiotis Karras, Qian Xiao, and Stéphane Bressan. "Sensitive Label Privacy Protection on Social Network Data." In *SSDBM*, pp. 562-571. 2012.

[24] Emelda, C., and R. Jaya. "Distributed Data Anonymization with Hiding Sensitive Node Labels." (2014).

[25] Campan, Alina, Yasmeen Alufaisan, Traian Marius Truta, and T. Richardson. "Preserving Communities in Anonymized Social Networks." *Trans. Data Privacy* 8, no. 1 (2015): 55-87.