SACK: Anonymization of Social Networks by Clustering of K-edge-connected Subgraphs

Fatemeh Heidari Soureshjani
Department of computer engineering, Payame Noor University, Po Box 19395-3697 Tehran, Iran

Arash Ghorbannia Delavar
Department of computer engineering, Payame Noor University, Po Box 19395-3697 Tehran, Iran

Fatemeh Rashidi
Department of computer engineering, Payame Noor University, Po Box 19395-3697 Tehran, Iran

ABSTRACT
In this paper, a method for anonymization of social networks by clustering of k-edge-connected subgraphs (SACK) is presented. Previous anonymization algorithms do not consider distribution of nodes in social network graph according to their attributes. SACk tries to focus on this aspect that probability of existence of an edge between two nodes is related to their attributes and this leads to a graph with connected subgraphs. Using connected subgraphs in anonymization process this method obtains better experimental results both in data quality and time. Sequential clustering is used for anonymization using k-edge connected subgraphs for starting step. Sequential clustering is a greedy algorithm and results are dependent on starting point.

Keywords
K-Anonymity, Social Networks, Privacy, Clustering, Information loss.

1. INTRODUCTION
By ever increasing spread of social networks, a huge amount of data is collected from individuals and their relationships. These data are valuable resources for researchers in different areas including social psychology, sociology, statistics and market research.[14] However they can be a threat for privacy of individuals, the owners who data is about them. So there is an essential need for data anonymization before its release. Social networks almost are new data structures. But these privacy concerns have been considered in traditional datasets, where the data could be simple relational tables, including sensitive information about individuals. Social networks have a more complex data structure, which contains some structural data, in addition to descriptive data. If consider a social network as a graph, individuals are represented as nodes of graph. Each node has some descriptive data like age, gender, race, country, majority, and the edges represent the relationships between them. Most Social network privacy methods are inspired from traditional ones like K-anonymity, which is a widely used privacy model for data anonymization. K-anonymity was first presented in [1], [2] and is used in different areas, which analyze huge amounts of data for revealing the hidden knowledge, like data mining[15], [16], [17], [18]. This anonymity model uses the concept of quasi identifiers, defined as subset of attributes, which can be used in linkage with other data sets to reidentify sensitive private information of individuals. The idea of this anonymization model is to generalize or suppress values of quasi identifiers in a way that each record in data set cannot be distinguished from at least k-1 other records (in this case study, a node with its descriptive and structural information). In other words if there is a combination of values of quasi identifiers in dataset, it must occur at least k times. To use k-anonymity model for social networks, new definitions are needed to make it compatible with graph data structure. Recently some methods have been presented for k–anonymity of social network which can be noticed by their main ideas. the first category are methods which use edge addition/deletion or switching edges of graph to prevent adversaries from identifying individuals with their knowledge about the structure of graph [4], [5], [6], [7], [8], [9]. These methods make changes in graph structure and the released data is different from original data. In the second category, data saves its original structure, but nodes are clustered and then each cluster is replaced with a super node, which will have all information of its contained nodes, both structural and descriptive information [9], [10], [11]. This study falls into second category, and focuses on the case of anonymization of social networks by clustering. These privacy preservation techniques are almost new, and try to find a clustering of nodes which minimizes the information loss measure. These methods present better results by improvements of clustering algorithm. Present study tries to in clustering process, consider distribution of nodes in social network graph according to their attributes and focuses on this aspect that probability of existence of an edge between two nodes is related to their attributes, such as age, country, etc. This leads to a graph with connected components. For this purpose the concept k-edge-connected subgraphs is used, which can define how to use this connected components in clustering process, That will result in better clustering of nodes with purpose of decreasing information loss.

2. RELATED WORK
K-anonymity of social networks by clustering was first considered by Zheleva and Geetor [11]. They presented the problem of sensitive relationships in social networks, and to address the problem, they used the concept link re-identification. Also they used a two-step anonymization method, first anonymization of descriptive information, without any attention to structural information. Then they presented five ways to anonymize the relationships. One of them is cluster-edge anonymization, which uses the aspect of anonymization of network by clustering.

The first anonymization algorithm that considers both descriptive and structural information at the same time was SaNGreeA and presented by Campan and Truta [9]. SaNGreeA starts clustering by selecting one node as first cluster, and continues adding nodes to this cluster till its size reaches k, then builds another cluster. Each time a node is added to current cluster which minimizes information loss.

Another algorithm presented by Tassa and Cohen [12], is sequential clustering. This algorithm builds all clusters at the same time, and in each iteration, moves one node between clusters, in a way that decrease information loss. Algorithm would stops when no such moves are available.

3. Anonymization Of Social Networks By Clustering Of K-Edge Connected Subgraphs

A social network SN is considered as a simple undirected graph G= (V, E), where $V = (v_1, \dots, v_n)$ is the set of nodes, and $E \subseteq \binom{V}{2}$ is the set of edges. Nodes represent individuals and edges represent their relations. Each node is described by a set of attributes and some of them are identifier attributes. These identifier attributes are removed from published data. But there is a set of attributes like zip code, gender, etc. called quasi identifiers, which can be used in linkage with other tables, to identify individuals. If some combinations of values of quasi identifiers be unique or rare, the adversary can determine the identity of related individuals.

The main goal of all anonymization algorithms is to reduce amount of information loss, within an acceptable time. The information loss in clustering based methods is result of two types of generalization: generalization of descriptive data, and generalization of structural data. Focus of this study is on reducing descriptive information loss.

Definition 1. Let $QI = (Q_1, \dots, Q_m)$ be the set of quasi identifiers, and for each Q_i , R_i be its domain , the descriptive data of SN will be $D = \{D_1, \dots, D_n\}$, where $D_i \in \{R_1 \times \dots \times R_m\}$.

Now, assume $Cl = (cl_1, \dots, cl_s)$ as a clustering of nodes in SN.

Definition 2. Let $D_1, \dots D_t$ be nodes of cl_i, the *similarity* of them on *j*th attribute will be defined as:

$$\boldsymbol{sim}(cl_i(j)) = \begin{cases} 1 - \dfrac{size\left(\left[\min\left(D_1(j), \dots, D_t(j)\right),\ \max\left(D_1(j), \dots, D_t(j)\right)\right]\right)}{size\left(\left[\min(R_j),\ \max(R_j)\right]\right)}, \\ \quad if\ j'th\ attribute\ is\ numerical \\ 1 - \dfrac{H(\wedge G(D_1(j), \dots, D_t(j)))}{H(T_j)}, \\ \quad if\ j'th\ attribute\ is\ categorical \end{cases}$$

Where size $size([a, b])$ is b – a, and $(\wedge T)$, is height of the tree rooted in T.

Accordingly, if *all* individuals have the same value on the *j*th attribute, the similarity will be 1. Similarity will be used for evaluation of information loss.

At this stage probability of existence of an edge between two nodes as is *introduced* follows.

Definition 3. Let D_u and D_v be two nodes in cl_i, and e_{uv} be the edge connecting them, the *Probability* of existence of e_{uv} will be:

$$p(e_{uv}) = \frac{1}{m}\sum_{j=1}^{m} sim(D_u(k), D_v(k))$$

By this definition , connectivity of a group of nodes can be defined as a ratio of number of edges between them, to all possible edges.

Definition 4. Let $D_1, \dots D_t$ be nodes in cl_i, connectivity between them will be:

$$connectivity(cl_i) = \frac{\sum_{1<u,v<t} p(e_{uv})}{\frac{t(t-1)}{2}} = \frac{\frac{1}{m}\sum_{1<u,v<t}\sum_{j=1}^{m} sim(cl_i(k))}{\frac{t(t-1)}{2}}$$

It is clear that connectivity of a cluster cl_i, is directly related to the similarity of its nodes. On the other hand, information loss resulting from anonymization by generalization of all nodes included in a cluster to a single node (a super node) and that is combination of structural and descriptive information loss:

$$IL = \alpha IL_D + \beta IL_S$$

Which IL_D and IL_S are respectively descriptive and structural information loss, and are weighted by α and , $\alpha, \beta \in [0,1]$, $\alpha + \beta = 1$. These weights can change due to the importance of descriptive or structural data quality preservation.

These information loss definitions are presented by Campan and Truta [9].

This paper focuses on the case of decreasing descriptive information loss, which defined as follows:

Definition 5. Let cl_i be a cluster, amount of descriptive information loss will be:

$$IL_D\left(G(cl_i)\right) = \begin{cases} \dfrac{1}{m_1}\displaystyle\sum_{j=1}^{m_1} \dfrac{size\left(\left[\min\left(D_1(j), \dots, D_t(j)\right),\ \max\left(D_1(j), \dots, D_t(j)\right)\right]\right)}{size\left(\left[\min(R_j),\ \max(R_j)\right]\right)}, \\ \quad if\ j'th\ attribute\ is\ numerical \\ \dfrac{1}{m_2}\displaystyle\sum_{j}^{m_2} \dfrac{H(\wedge G(D_1(j), \dots, D_t(j)))}{H(T_j)}, \\ \quad if\ j'th\ attribute\ is\ categorical \end{cases}$$

$$IL_D\left(G(cl_i)\right) = 1 - \frac{1}{m}\sum_{j=1}^{m} \boldsymbol{sim}(cl_i(j))$$

A relationship can be seen between connectivity of nodes included in a cluster, and similarity between them on their attributes. The connectivity of nodes in a cluster is directly depended to the similarity between them, and (according to the range of IL and sim, $IL, sim \in [0,1]$), has an inverse relationship with information loss. So starting clustering process by selecting nodes of initial clusters from connected subgrpahs, instead of random partition of nodes, the information loss will decrease.

To achieve this, before clustering process, find connected subgraphs, and then start clustering by focus on these subgraphs. one of the best algorithms to find these subgraphs is presented by zhou and liu [13]. A k-edge-connected graph is a connected graph that cannot be disconnected by removing less than k edges [13]. This is one of the best structures for extracting connected subgraphs. Fig 1 shows the steps of proposed method, named SACK, for anonymization of social network by clustering of k-edge-connected subgraphs.

Fig 2 shows graph of a sample social network SN, which has two 2-edge-connected subgraphs. Figs 3 to 6 illustrate the process of anonymization by SACK, and sequential clustering

by random partition of nodes. Here each node is assumed to be described by two attributes age and country, which for simplicity J is used for Japan and F for France. An anonymized version of network SN, is shown a graph super nodes(b).

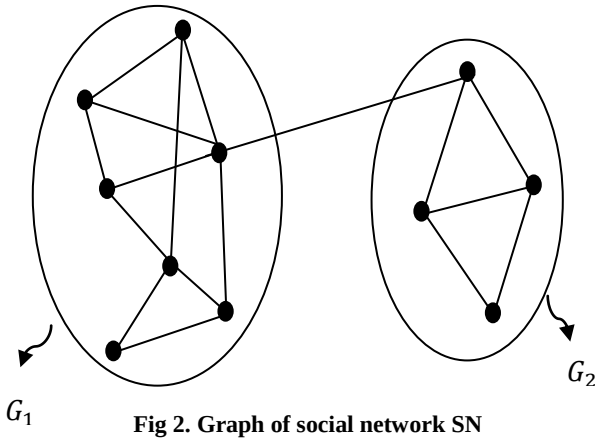


Fig 2. Graph of social network SN

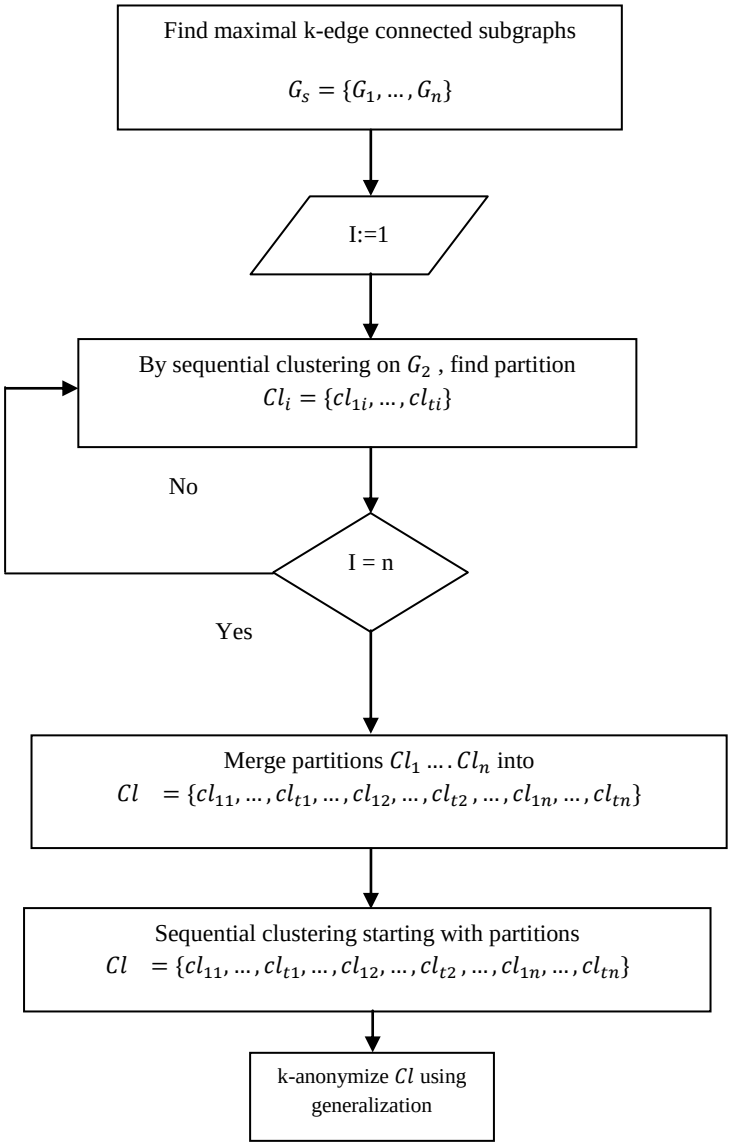


Fig 1. Flowchart of SACK

Fig 3 shows random partition of nodes that is used in sequential clustering. As previously mentioned nodes contained in a k-edge-connected subgraph have similar values of attributes, this aspect of social network can be seen in this sample. And it is obvious that a random partition of nodes puts nodes by different values of attributes in a cluster. Fig 4 shows clusters after one iteration, which one node moves between clusters to reduce information loss, and these iterations continue till no more moves are available, i.e. Relocation of nodes to any other cluster does not reduce information loss. In this way, in the best situation, sequential clustering by a random partition of nodes, after some iterations, leads to the state shown in fig 5, the state SACK starts process of clustering of k-edge connected subgraphs. Fig 6 shows the final clustering and 3-anonymized version of SN.

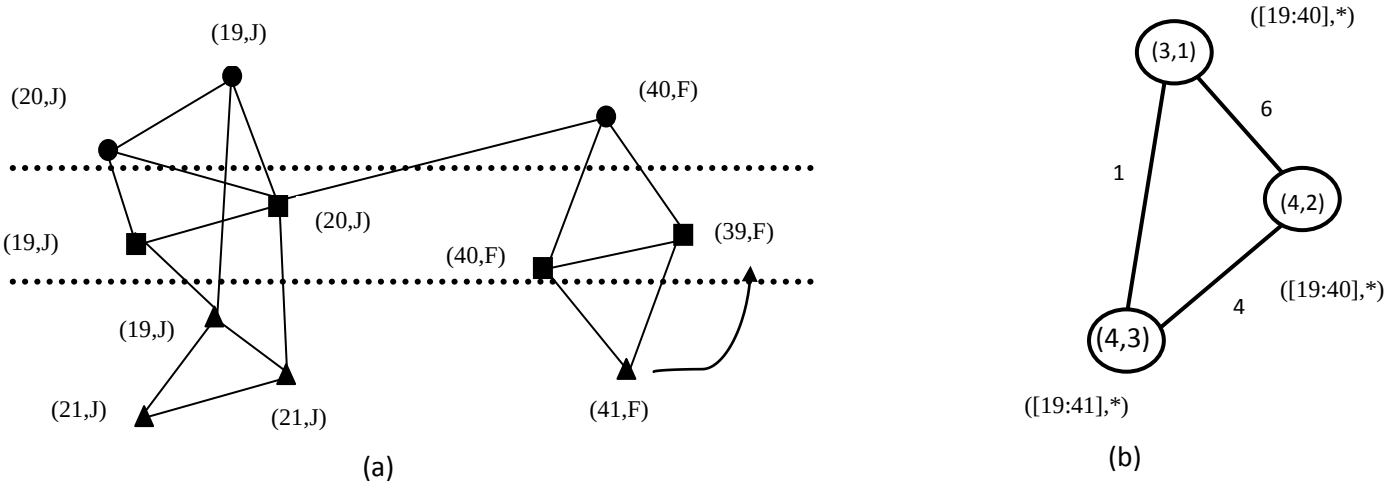


Fig 3. A network SN (a) and a corresponding clustering by random partition of nodes(b)

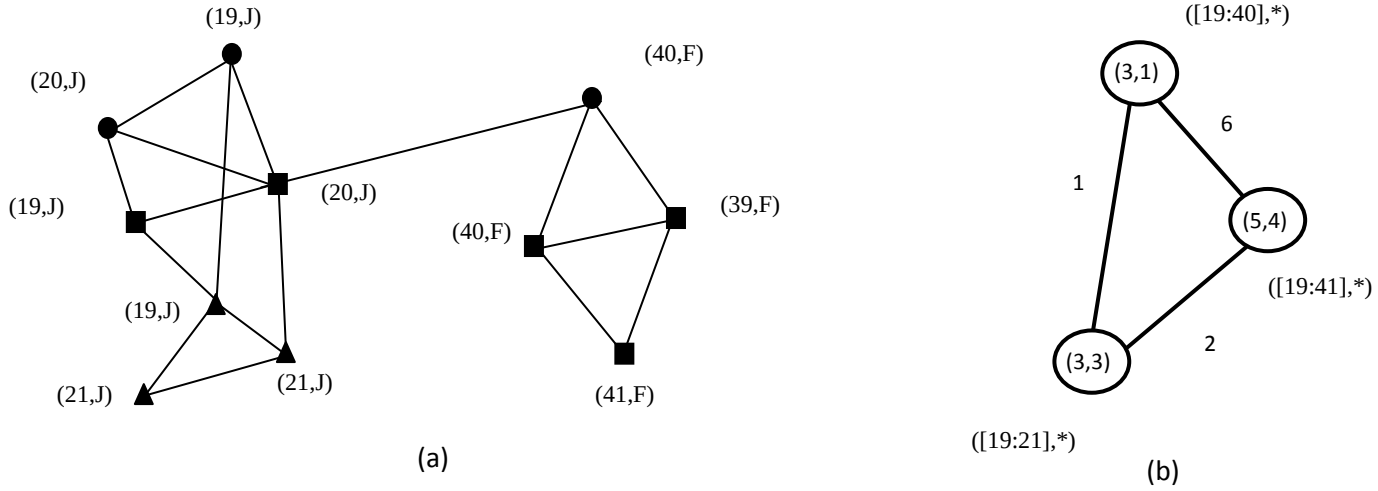


Fig 4. Network SN after one iteration (a) and its corresponding clustering (b)

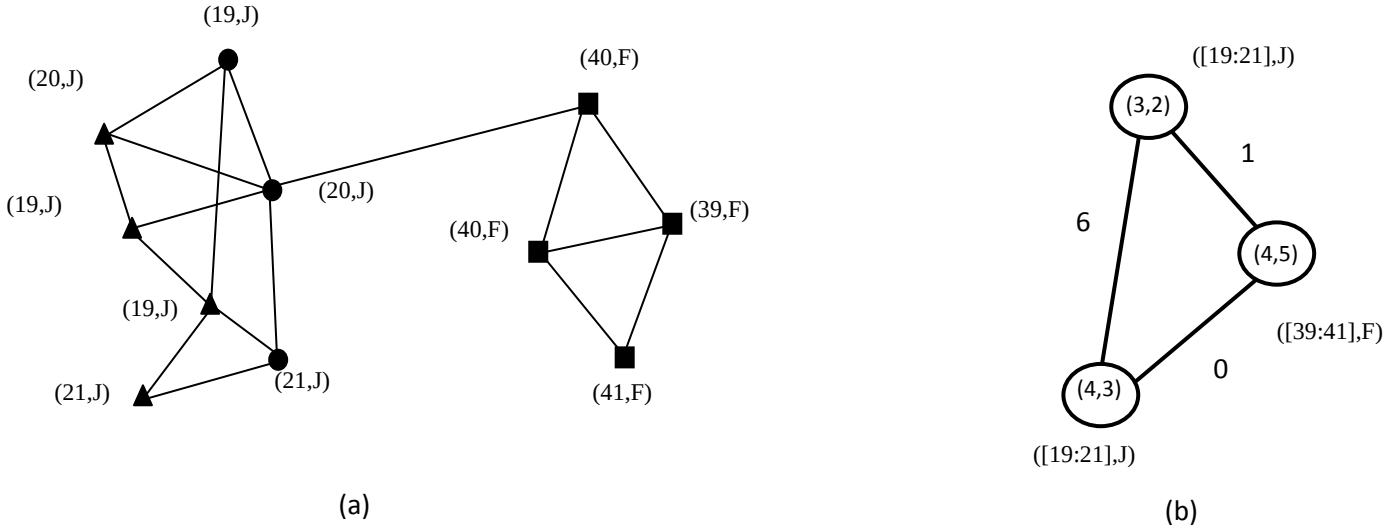


Fig 5. Network SN after some iteration (a) and its corresponding clustering (b)

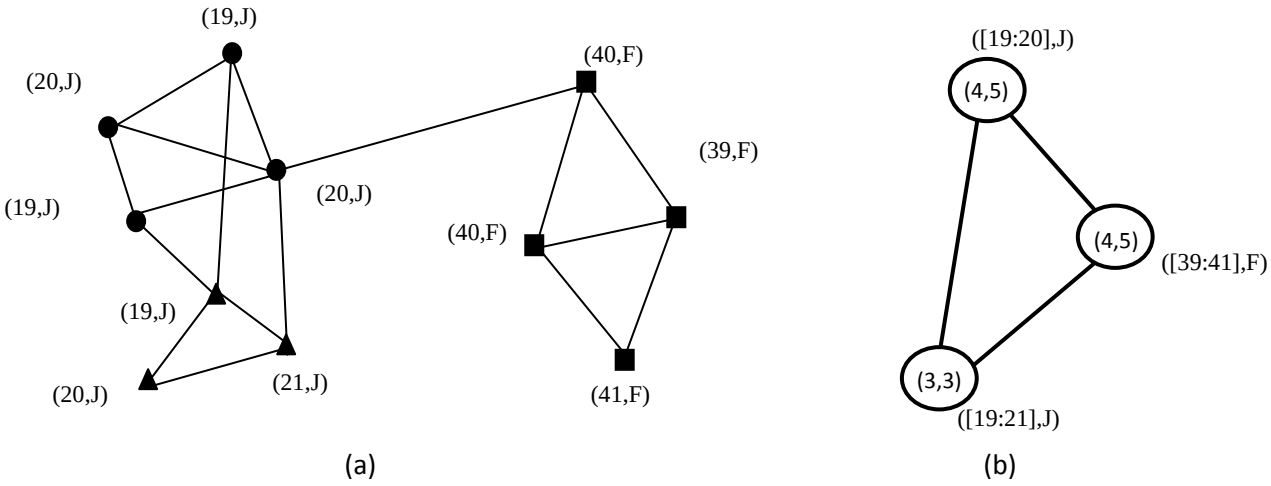


Fig 6. 3-anonymized version network SN

In this case, in samples with higher graph size, significant decrease of information loss is obtained, while there is no significant difference in amount of information loss in small samples. Due to the greedy nature of clustering algorithms, the starting point of clustering process has a significant impact on the results, and a random partition of nodes may lead to a local minima. The SACK algorithm is shown in fig 7.

Algorithm *SACK*

Input: social network SN,graph G , k , k

Output: Cl, a clustering of nodes in SN

1. Find maximal k-edge connected subgraphs

$$G = \{G_1, \dots, G_n\}$$

2. for i=1...n

3. Sequential clustering on G_2 *, find partition*

$$Cl_i = \{cl_{1i}, \dots, cl_{ti}\}$$

4. end loop

5. Merge partitions into $Cl_1 \dots Cl_n$ *into* $Cl = \{cl_{11}, \dots, cl_{t1}, \dots, cl_{12}, \dots, cl_{t2}, \dots, cl_{1n}, \dots, cl_{tn}\}$

6. *Sequential clustering on Cl , starting with partitions* $Cl = \{cl_{11}, \dots, cl_{t1}, \dots, cl_{12}, \dots, cl_{t2}, \dots, cl_{1n}, \dots, cl_{tn}\}$

7. RETURN *Cl*

Fig 7. SACK algorithm

Here k is used for k-anonymity a k for k-edge-connected graph, to avoid ambiguity.

4. EVALUATION RESULT

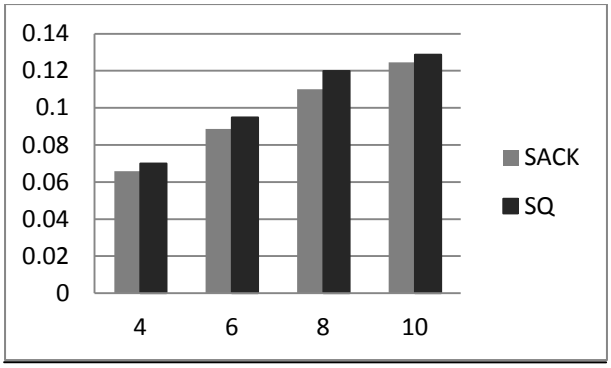
In this section, the experimental results of SACK against sequential clustering (SQ) are presented. Different datasets are used to compare both data quality and runtime. The descriptive data is extracted from adult dataset from the UC Irvine Machine Learning Repository, and the structural data is generated by the definitions of similarity and connectivity, mentioned in section 3. The attributes included from adult dataset are age, gender, education level, marital status, race, work class and nationality.

First on a dataset of 147 nodes, SACK and SQ are tested on different values of k, due to k-anonymity. Fig 8 illustrates the information loss results from both sequential clustering (SQ) and SACK, indicating significant improvements in SACK. On the other hand, fig 9 shows runtime results, which as could be expected, SACK has a smaller runtime than SQ.

Figure 10 shows the runtimes of the SACK and SQ, for *k* = 25, where size of the dataset ranges from n = 210 to n= 300.

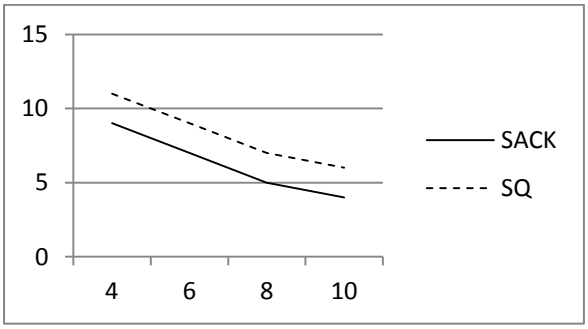
5. CONCLUSIONS

In this study, a method for anonymization of social networks by clustering of k-edge-connected subgraphs is proposed. This method tries to by improvement of starting point in clustering process, reduce both information loss and runtime. Previous clustering based anonymization algorithms do not consider distribution of nodes in social network graph according to similarity of descriptive information. These similarities lead to a graph with connected components. Current study used these connected components to improve starting point of clustering. Experimental results show that this method performed well both in data quality and runtime.



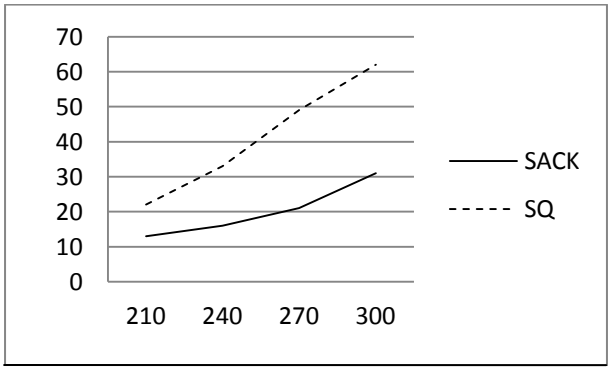
k	SACK	SQ
4	0.06	0.07
6	0.08	0.09
8	0.11	0.12
10	0.12	0.13

Fig 8. Information loss vs. k



k	SACK	SQ
4	9	11
6	7	9
8	5	7
10	4	6

Fig9. Time vs. k



n	SACK	SQ
210	13	22
240	16	33
270	21	49
300	31	62

Fig10. Time vs. N

6. REFERENCES

[1] P. Samarati. Protecting Respondents Identities in Microdata Release. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 13, No. 6, 1010–1027, 2001.

[2] L. Sweeney. K-Anonymity: A Model for Protecting Privacy. *nternational Journal on Uncertainty, Fuzziness, and Knowledge-based Systems*, Vol. 10, No. 5, 557–570, 2002.

[3] F. Bonchi, A. Gionis, and T. Tassa. Identity obfuscation in graphs through the information theoretic lens. In *ICDE*, pages 924–935, 2011.

[4] S. Hanhij¨arvi, G. Garriga, and K. Puolamaki. Randomization techniques for graphs. In *SDM*, pages 780–791, 2009.

[5] M. Hay, G. Miklau, D. Jensen, P. Weis, and S. Srivastava. Anonymizing social networks. *Uni. of Massachusetts Technical Report*, 07(19), 2007.

[6] X. Ying and X. Wu. Randomizing social networks: A spectrum preserving approach. In *SDM*, pages 739–750, 2008.

[7] X. Ying and X. Wu. Graph generation with prescribed feature constraints. In *SDM*, pages 966–977, 2009.

[8] X. Ying and X. Wu. On link privacy in randomizing social networks. In *PAKDD*, pages 28–39, 2009.

[9] A. Campan and T. M. Truta. Data and structural k-anonymity in social networks. In *PinKDD*, pages 33–54, 2008.

[10] M. Hay, G. Miklau, D. Jensen, D. F. Towsley, and P. Weis. Resisting structural re-identification in anonymized social networks. In *PVLDB*, pages 102–114, 2008.

[11] E. Zheleva and L. Getoor. Preserving the privacy of sensitive relationship in graph data. In *PinKDD*, pages 153–171, 2007.

[12] Tamir Tassa, Dror J. Cohen: Anonymization of Centralized and Distributed Social Networks by Sequential Clustering. IEEE Trans. Knowl. Data Eng. 25(2): 311-324, 2013.

[13] Rui Zhou, Chengfei Liu, Jeffrey Xu Yu: Finding Maximal k-Edge-Connected Subgraphs from a Large Graph. In EDBT '12, Pages 480-491, 2012.

[14] Freeman L. The Development of Social Network Analysis. Vancouver: Empirical Press, 2006.

[15] Aggarwal G, Feder T, Kenthapadi K, Motwani R, Panigrahy R, Thomas D, Zhu A. Anonymizing Tables. ICDT, pp:. 246-258, 2005.

[16] Bayardo R. J., Agrawal R. (2005). Data privacy through optimal k-anonymization. In Proc. of the International Conference on Data Engineering (ICDE), Tokyo, Japan, 2005.

[17] Chawla S, Dwork C, Mcsherry F, Smith A, Wee H. Toward privacy in public databases. In Theory of Cryptography Conference (TCC), 2005.

[18] Ciriani V, Vimercati S, Foresti, S, Samarati P. k-Anonymous data mining:A survey.Book:Privacy-preserving data mining. springer US, pp: 105–136, 2008.