ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Graph perturbing Method for Privacy Preserving against Neighborhood-Pair Attacks

K.Venkata Ramana
Department of CS & SE,
Andhra University College of Engineering (A),
Visakhapatnam, AP.

Anusha Piratla
Department of CS & SE,
Andhra University College of Engineering (A),
Visakhapatnam, AP.

ABSTRACT:

A lot of privacy models and anonymization techniques have been developed to prevent re-identification of nodes. A PKDLD anonymity model has been developed for providing individual privacy. Even though privacy is provided, re-identification of the vertex is the major problem in social network data publishing.

A new attack known as neighborhood pair attack had been proposed which uses the structural information of the node. In this paper, a solution to the neighborhood pair attack is demonstrated by means of random graph perturbation technique. Thus, re-identification of an individual node is prevented. Therefore the probability of re-identifying the node is reduced even when the data is published.

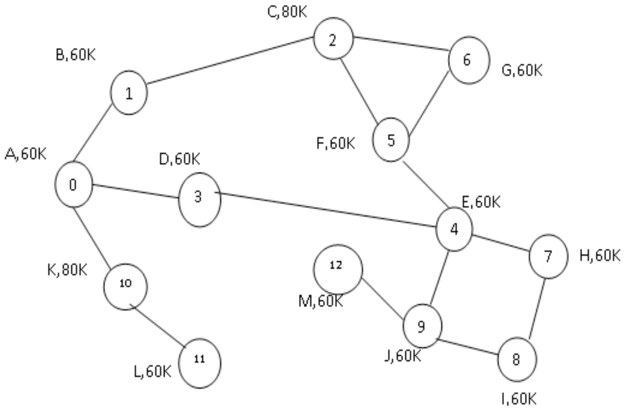
Keywords:

Analysis of social network, Privacy Preserving.

1.INTRODUCTION:

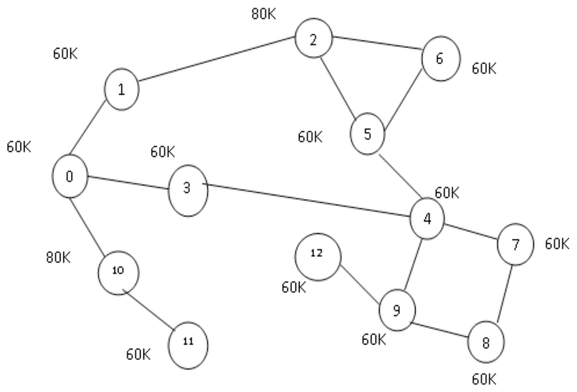
Whenever data is published in social networks, preserving the privacy of an individual become a major problem. Nowadays, a variety of anonymization techniques have been developed to protect the data in social network. Some of these approaches are naive anonymization, k degree anonymity, l-diversity etc to preserve the privacy of individuals.

Therefore privacy preserving of publishing social network data is a serious concern. Each vertex in the below social network is represented as person and the edges represents the relation between them as shown in fig(a).



Fig(a): Social Network

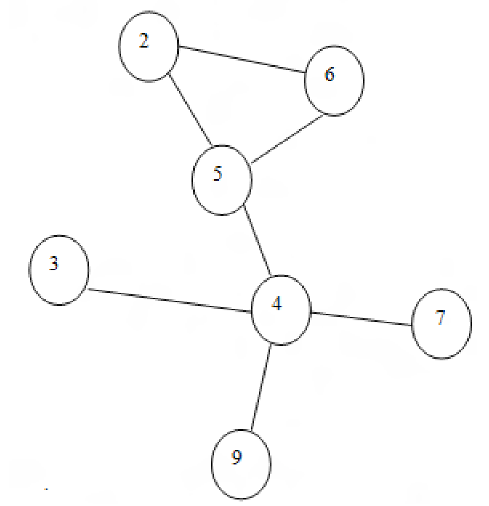
Now to preserve the piracy, we remove the identifying attributes and publish the data as shown in fig(b)



Fig(b): Naive Anonymized Network

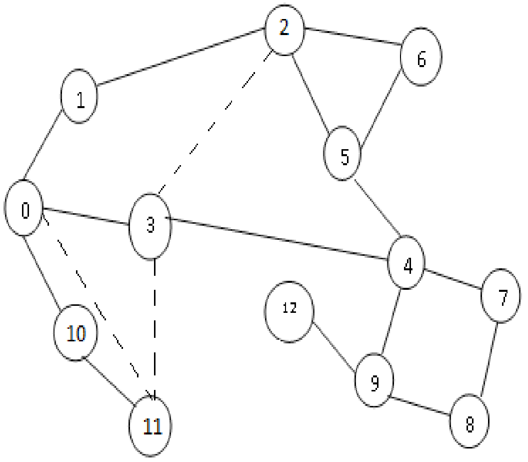
Now, if an adversary knows the neighborhood structure of a pair of connected vertices, then re-identification of target node is possible. For eg, if the adversary node knows that the degree of node 4 is four and one of the neighbors of node 4 has degree 3 and including node 4 and two of its neighbors are directly connected to each other. So, vertex 4 can easily be re-identified since no other vertex matches the above criteria.

The neighborhood-pair structure of node 4 is shown below:



Fig(c): Structural Information of node 4

Therefore, In order to protect the privacy, we prevent a graph perturbation algorithm by adding noise edges between (0,1),(3,11),(2,3) as shown below. Now we have three similar nodes with the same structural information i.e nodes:4,2,0.Therefore adversary can identify the target node with a probability greater than 1/3. The perturbed graph is shown in Fig(d)



Fig(d): Perturbed Social Network Graph

2. PROBLEM DEFINITION:

A. preliminaries:

We model a social network as a graph G=(V,E) where v={v1,v2,v3...vn} is a set of nodes and E is a set of unlabelled, undirected edges, E C (VXV). In Social Network each node corresponds to an individual and an edge represents a relationship between two individuals.

Here, the given graph is 3 anonymous means atleast (k-1) i.e 2 nodes have same degree. We take the scaling factor as (k-1) i.e 2.It means atleast two nodes in the graph must be isomorphically equivalent such that the structure around the nodes remains same. We can achieve this by anonymizing a social network by applying noise edges to preserve the structure of the uniquely identifiable nodes. This technique helps in reduction of re-identification of targeted nodes. The main aim is to resist attack from the adversary.

DEFINITION 1 (Randomized Social Network): An anonymizing technique in which random edge insertions and deletions are done.

DEFINITION 2 (Naive Anonymization) : To protect the node from uniquely identified , node names are replaced by synthetic identifiers. This is known as Naive Anonymization.

DEFINITION 3 (Structural Equivalence) : Structural Equivalence refers to the extent to which two nodes are connected to the same other i.e., have the same social environments. It is often hypothesized that structurally equivalent nodes will be similar in other ways as well, such as in attitudes, behaviors or performance.

DEFINITION 4 (Relative Equivalence) : If two nodes a, b are relatively equivalent if Hi(a)= Hi(b) and is denoted by a=Hi(b)

GRAPH INFORMATION:

H0(x) returns the label of a node.

H1(x) returns degree of a node.

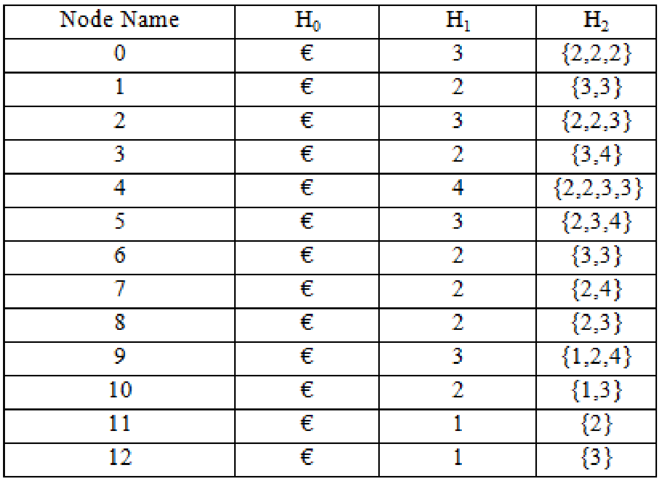
H2(x) returns multiset of each neighbors degree.

Hi(x)={Hi-1(z1), Hi-1(z2),......, Hi-1(zm)}

where z1,z2,...,zm are neighbors of node x.

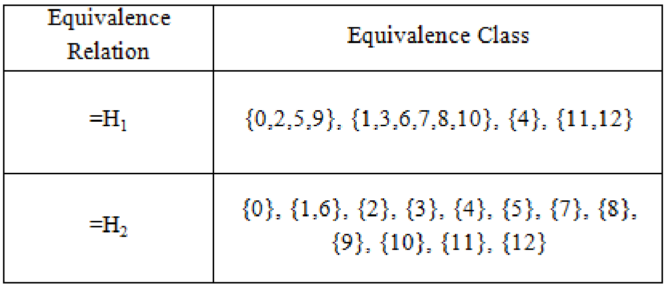
Table(i) shows the graph information table and Table(ii) shows the equivalence of the graph shown in fig(d).

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal



Node Name	H_0	H_1	H_2
0	€	3	{2,2,2}
1	€	2	{3,3}
2	€	3	{2,2,3}
3	€	2	{3,4}
4	€	4	{2,2,3,3}
5	€	3	{2,3,4}
6	€	2	{3,3}
7	€	2	{2,4}
8	€	2	{2,3}
9	€	3	{1,2,4}
10	€	2	{1,3}
11	€	1	{2}
12	€	1	{3}

Table(i): Graph Information Table



Equivalence Relation	Equivalence Class
$=H_1$	{0,2,5,9}, {1,3,6,7,8,10}, {4}, {11,12}
$=H_2$	{0}, {1,6}, {2}, {3}, {4}, {5}, {7}, {8}, {9}, {10}, {11}, {12}

Table(ii): Equivalence Table

3. RANDOM GRAPH PERTURBATION ALGORITHM:

Input: Personalized Social Network G=(V,E) in form of Adjacency Matrix and Anonymization parameter K

Output: Perturbed Graph G'

Method:

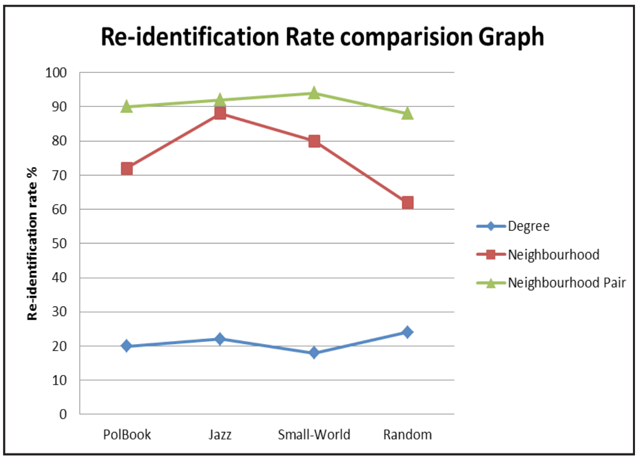
```
// Calculation of degree
1.for all (nodes vi in V) then
2.Calculate H1 of vi and place in H1[i] array
3.i++
4.repeat steps from 1 to 3 until V is empty
// calculation of multiset of each neighbors degree.
5.for all (nodes vi in V)
6.Calculate multiset of neighbors from ADJ[i][j] and place in an array H2i[H1[j]]
7.increment i value
8.repeat steps from 5 to 7 until V is empty.
// perturbation of graph by means of random edge deletions and insertions.
9.for all (nodes in vi) then
10.Compare H2i[ ] matrix for every vi
11.If(H2i[ ] does not match with any other H2[j] ) then
12.add or delete edges between the nodes in ADJ[i] [j] such that atleast K  nodes are isomorphically eqivalent
13.end if loop
14.end for loop.
```

4. Experiments:
Dataset Used:

We use both real datasets and synthetic datasets. Random graphs are generated using sunthetic datasets for experimental analysis. Random graphs: Graphs with nodes randomly connected to each other with probability p are called Random graphs . Given the number of nodes n and the parameter p, a random graph is generated by creating an edge between each pair of nodes u and v with probability p. For the real datasets, we use the PolBooks, Jazz, Small World, Random PolBooks: A network of books sold by an online store where the edges between books represent the purchase frequency of the same buyers. Jazz: A network of jazz musicians who collaborate in different bands. The vertices represent the band and edges represent the musicians in common. Small-World: A type of graph in which most vertices can be reached from every other vertex by a small number of hops. Random: a synthetic random network where the vertices in this network are randomly connected.

5. Results and Analysis:

The graph in Fig(e) compares the re-identification rate among the three types of structural information over the five different datasets:

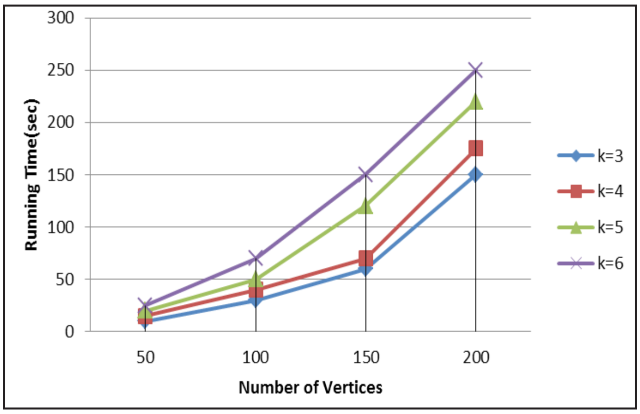


Fig(e):The Re-identification Rate Comparision Graph

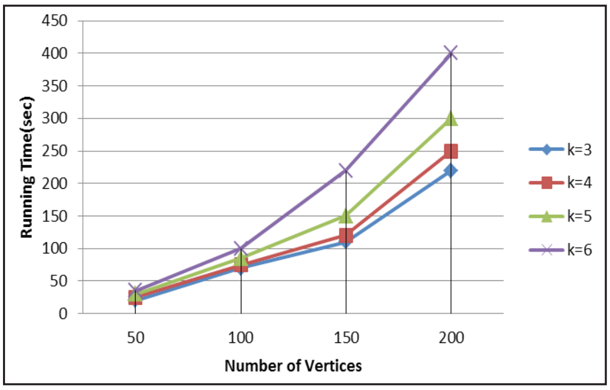
The percentage represents the number of vertices the dataset that are exposed to re-identification attack using the three types of graph structural information. Therefore, this graph only shows the rate of vertices that definitely re-identified. The degree attack identifies 20–30 % of the users to be re-identified. The neighbourhood structural information gives more than double of the reidentification risk. Through all the datasets, neighbourhood attack identifies 60% of the users. However, it is evident that the reidentification rate using the neighbourhood-pair scored the highest in all the datasets.Therefore, using our graph perturbation algorithm we reduce the re-identification of the nodes and an analysis is done on the runtime and cost on various synthetic datasets.

Analysis of runtime on various synthetic datasets:

The runtime on various synthetic data sets with respect to different K values is shown in Fig(f) and Fig(g). The runtime increases when the average vertex degree increases, since the network becomes denser. Moreover, the larger the k, the longer the runtime since more neighborhoods in a group need to be anonymized.



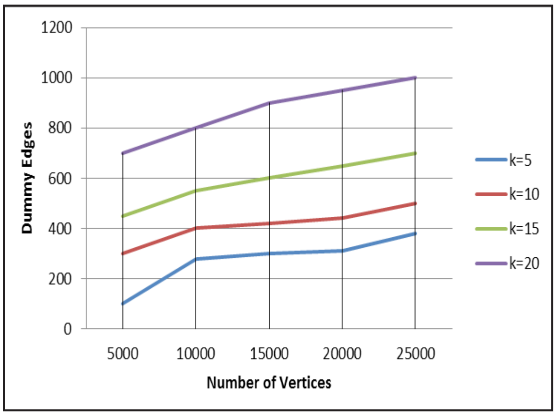
Fig(f) Average Vertex Degree = 3



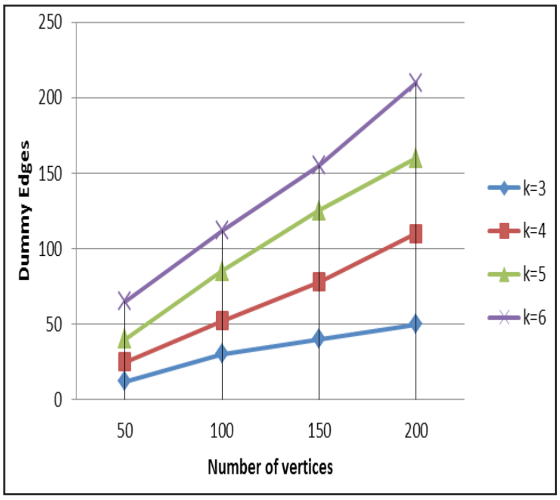
Fig(g) Average Vertex Degree = 5

Anonymization cost on various Synthetic Datasets:

The anonymization cost in the number of edges added on various synthetic data sets with respect to different k values is shown in fig(h) and Fig(i). First, when the number of edges increases, the anonymization cost increases. Second, when k increases, the anonymization cost also increases, because more neighbourhood-pairs are needed to be anonymized in a group. Last, when the average number of vertex degree increases, the anonymization cost increases, too. In a denser network, the neighbourhood-pairs are more diverse and more number of edges are needed to anonymize different neighbourhood-pairs.



Fig(h) Average Vertex Degree=3



Fig(i) Average Vertex Degree= 5

6. Related Work:

As the technology is advancing, it is possible to collect data of the individuals and the relationship between them.

Protecting the privacy of individuals is the main concern. Privacy is typically protected by anonymizing . A Novel Anonymizing technique has been proposed by Hay [1] which is based on perturbing the network. Perturbation is a promising technique for enhancing anonymity. Our goal is to enable the useful analysis of social network data while protecting the privacy of individuals. Many Anonymization techniques had been introduced to preserve privacy such as Link Privacy [2], k-anonymity[3], l-diversity[4].Most of the existing methods did not cater for the individuals' personalized privacy requirements and did not take full advantage of distributed characteristics of the social network nodes. Motivated by this, Jia Jiao specify three types of privacy attributes for various individuals and develop a personalized k-degree-l diversity (PKDLD) anonymity model [5]. An essential type of privacy attack has been introduced by Jein Pei known as Nieghborhood Attack[6] in which the adversary knows the background knowledge of the target node. Pei and Zhou proposed a k-anonymity model to prevent this attack. In this paper, solution to the Neighborhood Pair attack[7] has been presented in which the attack is preserved by means of Graph Perturbation Technique which prevents the target node from re- identification.

7. CONCLUSION:

In this paper an initiative is taken to combat neighbourhood-pair attack. We modeled the problem and developed an approach that fights neighbourhood-pair attack. A study is conducted on both real data set and synthetic data sets strongly indicate that neighbourhood-pair attacks are real in practice, and our method can be done in practice. As data is increasing day by day, serious efforts are necessary in future because social network data is much more convoluted than relational data, privacy preserving in social networks is much more challenging and needs many serious efforts in the future.

Only neighbourhood-pair attack is handled in this paper. However, It will be very interesting and if d-neighborhoods (d > 2) are protected. However, this will introduce a serious instigation in computation. As d increases, the neighborhood size increases exponentially. One of the major problem that will be faced is in conducting Isomorphism tests and anonymization of large neighborhoods in a network becomes very difficult. Privacy may not be completely preserved eventhough the the network is K-anonymous.

If an adversary can identify a victim in a group of vertices anonymized in a group, but all are associated with some sensitive information, then the adversary still can know that sensitive attribute of the victim. Therefore, we need to anonymize the network in such a way that both the sensitive attributes as well as node re-identification must be preserved.

REFERENCES:

[1]Hay, M., Miklau, G., Jensen, D., Weis, P., Srivastava, S :" Anonymizing social networks". Computer Science Department Faculty Publication Series, pp. 180, 2007.

[2]Aleksandra Korolova, Rajeev Motwani, Shubha U. Nabar, Ying Xu:" Link Privacy in Social Networks". In Proceedings of the 17th ACM conference on Information and knowledge management, pp. 289-198, 2008.

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

[4]Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M.: l-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD) 1(1), 3 (2007).

[5]Jia Jiao, Peng Liu, and Xianxian Li.: "A Personalized Privacy Preserving method for Publishing Social Network Data". In proceedings of 11th Annual Conference, TAMC 2014-Springer, pp. 141-157, 2014.

[6]Zhou, B., Pei, J.:" Preserving privacy in social networks against neighborhood attacks". In proceedings of IEEE 24th International Conference on Data Engineering, ICDE 2008, pp. 506–515. IEEE, 2008.

[7]Mohd Izuan Hafez Ninggal and Jemal H. Abawajy.: "Neighborhood-Pair attack in Social Network Data Publishing". In proceedings of 10th International Conference, MOBIQUITOUS 2013-Springer, pp. 726-731, 2014.