International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Implementation Details of Anonymization of Sensitive Labels in Collaborative Data Publishing

Vandhana V[1]

[1]M.Tech Research Scholar
Department of Computer Science & Engineering
Sree Buddha College of Engineering, Alappuzha, Kerala, India

Abstract: ***Nowadays, no: of research and development in privacy preserving publishing of social networking data .The privacy is one of the important concerns for the social network. So, our aim is protect the privacy preserving publishing of social networking data in collaborative environment using the k-anonymity model, the set of nodes share the same attributes so, the sensitive information may gain access by the intruder. So, our aim is that how to protect the sensitive attributes of individuals and the structural information of social network data in collaborative environment. In this paper we study the existing structure anonymization mechanisms and defined a model called k-degree-l-diversity anonymity model, which takes into consideration the structural information and sensitive labels of individuals in collaborative data publishing. Collaborative data publishing means that different publishers publish their data independently for e.g. in hospitals, wish to publish anonymzed view of their data. We have large no: of dataset it's not anonymized so; the attacker should be hack the data. So, they protect the data's of different publishers from the hackers by using a new anonymization method called k-degree –l diversity model.***

Keywords: Anonymization, Collaboration, Social network, Sensitive attributes

1. Introduction

Usage of different social networking sites has increased frequently in the recent year. Due to this the attackers have an opportunity to access the useful information such as community growth, disease spreading in a particular region etc. The social networking sites must preserve the private information of the individuals and ensuring the privacy and utility. So, we discuss about the existing privacy models and graphic models. Social networks are represented in the form of graphs. Each node is called individuals of the social network and the corresponding private information of the individuals is stored as the entities of the node. Privacy is the one of the main challenging issue for publishing the social networking data. The private information of an individual is the age, email id, salary etc. They publishing the social networking data they protect the private information of the individuals. Our aim is that how to protect the structural information and the sensitive label of individuals in collaborative environment. Several types of anonymization techniques are used for the privacy preserving publishing of social data but they do not protect the private information of individuals in collaborative environment. The several approaches are Clustering, Edge editing, Graph modeling etc are used for the single source data publishing. In the Edge editing approach they modify the graph by adding or deleting edges. They protect the graph from the re-identification of nodes. The clustering based approach they clustering the nodes and edges. These clustering data's into no of groups and they anonymized into a super node. All the individuals' data's are hidden into that super node [9]. The different types of attacks are in the graph structure the one of them is structural attack. The structural attack means they attack the private information of the individuals. In this case the attacker could re-identify the information from the subgraph, nodes, edges etc [8].

Another type of attack is called social intersection attack [10]. In the social intersection attacks, the compromised users can identify the originators of shared contents. This type of attack is a graph anonymization problem, to avoid this problem the author proposed a new approach called the users with k-anonymity privacy. The privacy becomes guarantees by supplementing the social graph with latent edge.

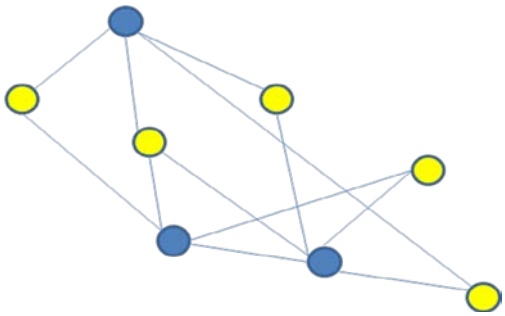

Figure 1: Data represented in the form of graph

The fig 1 shows the representation of data's in the social network in the form of a graph. Each node in the graph shows the individuals (blue) and corresponding sensitive and non sensitive labels (yellow) are connected to the node through the edges.

2. System Design

In (2013) introducing a paper called Protecting Sensitive Labels in Social Network Data Anonymization. In this paper more concerned about the privacy preserving publishing of social data. The social data should be anonymized and then publishing the dataset in the form of a graph. In this case the original graph is modified into the KDLD graph that is k-degree l-diversity graph [8]. Using the KDLD algorithm they protect the structural information and sensitive labels like age, email id, disease details etc of individuals in single

source data publishing. One of the disadvantages for this anonymized model is not defined how to protected the structural information and sensitive labels of individual in collaborative environment.

In this work, we define a k-degree-l-diversity anonymity model that considers the protection of structural information as well as sensitive labels of individuals in the collaborative environment. We accept data sets from widely known dataset collection from Internet. The dataset will be maintained to formulate clusters, their by checking degree calculation. Then we develop a new algorithm by adding noise nodes into the original graph with the consideration of introducing the least distortion to graph properties. Most importantly, we provide a rigorous analysis of the theoretical bounds on the number of noise nodes added and their impacts on an important graph property. For implementing the new system consist of no of steps like Add publishers, generalization, merging, display graph of merging dataset, anonymization then anonymized graph.

All paragraphs must be indented. All paragraphs must be justified, i.e. both left-justified and right-justified.

2.1 Add Publishers

In this module add the no of publishers. These publishers publish their data into the social networking sites e.g. for publishing data like survey of salary etc. In this step they collect the data's of the different publishers like the name of the publisher, profile details and upload their dataset of each of the publishers.

2.2 Generalization

In this module they generalise the uploaded dataset in this module merge the datasets of add publishers. Merging should be done on the basis of sensitive attributes. In merging steps first upload the dataset of first publisher then the second publisher and so own.

2.3 Integrate

In this module integration of dataset should be done. In the integration step node reducing should be done. The node redundancy avoided by comparing each of the rows of the dataset on the basis of sensitive attribute. If the two dataset contain same sensitive attribute they remove the one of the same attribute row from one dataset. Then update the edges of the removed node. These edges are added into the compare node. If no one with similar sensitive attribute only combined the datasets. After this step a new combined dataset should be generated. Then generate the graph of the publishers.

2.4 Display Graph

The graph generating software is used in this module. The software takes the input and processed the combined dataset and generates graph. It is based on force directed algorithm and here we used Graph#. This software contains many algorithms for the graph generating like tree, circle, and frequency based algorithms etc. The output is a single graph i.e., each publisher has its own graph generated. If the datasets are contains the common data they become eliminating that node from the one of the dataset and generated a single graph for all the publishers. After generation of the graph they should be saved into the out graph folder in the GML format.

2.5 Anonymization

In this module the dataset should be anonymized before publishing the social data of individual by publishing the different publishers. KDLD anonymized method is used for the data anonymization.In this case first upload the GML dataset and data management process should be done. The data management step the dataset is in CSV format. The details of the nodes- the id and the labels form the first few lines followed by edge description- the id and the end nodes forming the undirected edge. They convert the GML format into table format. This helps in easy conversion of the dataset to two tables- the node table and the edge table. The node table includes node id and the different labels. The edge table includes the edge id and the end nodes. The dataset is divided into no of groups which satisfies three constraints.

1. All the elements in the group have same degree.
2. At least size k
3. The groups have at least l distant label values. After that they perform the APL preserving algorithms. In the APL preserving find the average shortest path length between the nodes in the graph. The five step algorithm is performing in the groups and adding some of the noise node and edges in the graph for generating anonymized dataset.

2.6 Display Anonymized Graph

In the anonymization step we have to generate the anonymized dataset. This data set is used for the anonymized graph generation. In this modified graph should protect the private information of individuals from the hackers.

3. System Implementation

3.1 Home Page

This is the main home page. This contains all the steps that have to be done in this work.

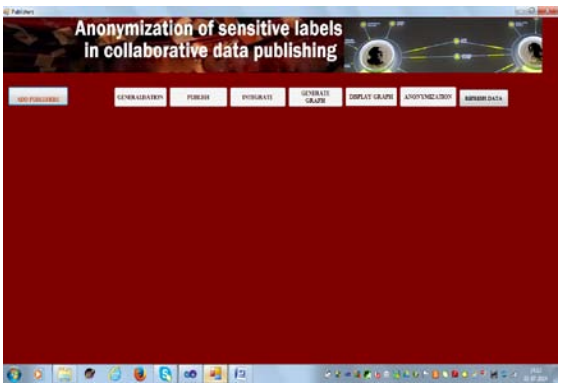

Figure 3.1: Home page

3.2 Add Publishers

In this fig shows the first module of this project like add publishes and each publisher's details stored into the database. The publisher's details like publisher name, profile and the dataset are saved. In this module we can save the any no: of publisher's details.

Figure 3.2: Add Publishers

3.3 Generalization

The fig 3.3 shows the second module. In the generalization module generalize the publisher's dataset. In this step all the publisher's datasets combined into a single dataset. The dataset is divided into two tables' node table and edge table. The node table contains the details of each node in the dataset. The edge table contains the edges of each node in the dataset.

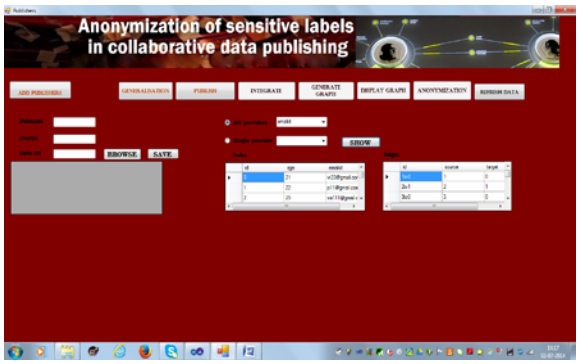

Figure 3.3: Generalization

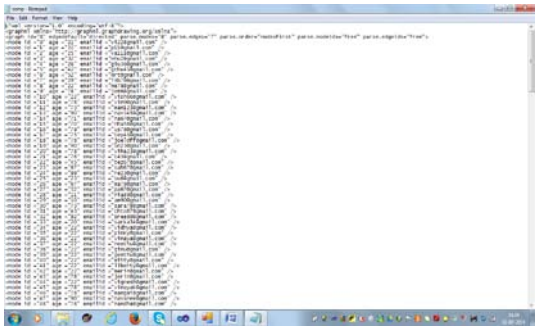

Figure 3.4: Combined dataset

3.4 Integrate

The fig 3.4 shows the integration of dataset. In this module shows the node reducing. Before integration they published the generalized dataset. This generalized dataset is used for the node reducing. The generalized dataset should contain the overlapping of data. The different publishers publish their data independently and they become overlapping. So, they reduce the overlapping data. Overlapping means that the redundant nodes are in the dataset. In this module they were reducing the redundant nodes from the new dataset. The node redundancy avoided by comparing each of the rows of the dataset on the basis of sensitive attribute. If the two dataset contain same sensitive attribute they remove the one of the same attribute row from one dataset. Then update the edges of the removed node. These edges are added into the compare node. If no one with similar sensitive attribute only combined the datasets. After this step a new combined dataset should be generated. Finally, this dataset is used for the generation of graph of the publishers.

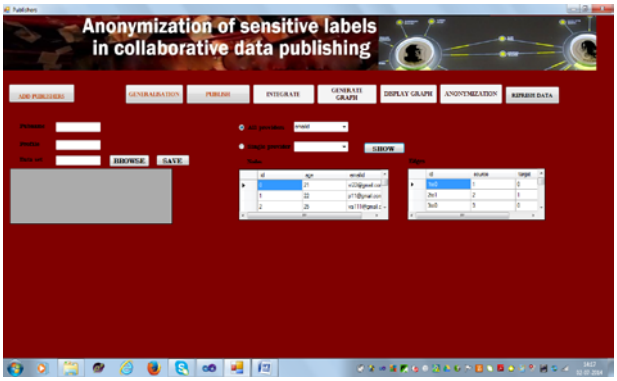

Figure 3.4: Integrate

3.5 Display Graph

The fig 3.5 shows the graph of publishers. In this module a graph generating software is used for generating the graph. In this case two publishers graph is shown in the figure. The two publishers have different data's so no elimination will happened. That's why a single graph is generated.

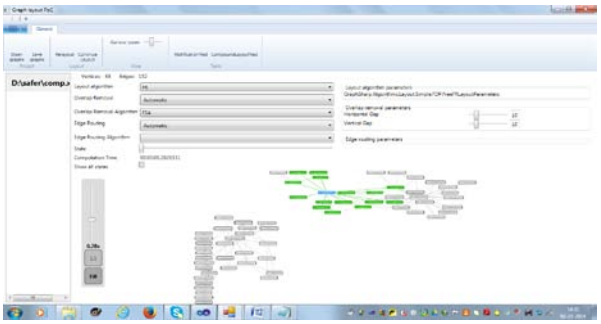

Figure 3.5: Graph of Publishers

In fig 3.6 shows the dataset in the GML format. The generation of publishers graph should be saved into a folder called out graph. The saved graph is GML format. This dataset is used for the anonymization.

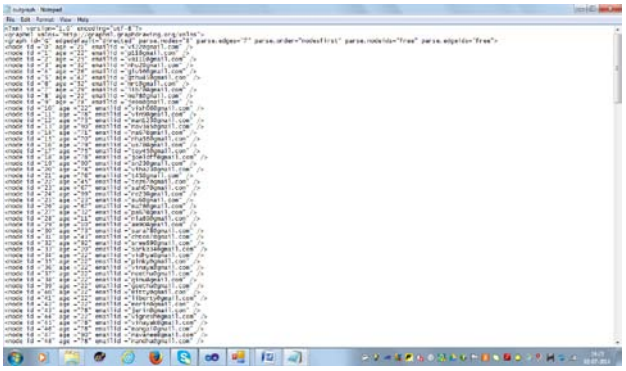

Figure 3.6: Graph in GML format

3.6 Anonymization

3.6.1 Step 1: Data Conversion

The details of the nodes- the id and the labels form the first few lines followed by edge description- the id and the end nodes forming the undirected edge. They convert the GML format into table format shows in fig 3.7. This helps in easy conversion of the dataset to two tables- the node table and the edge table. The node table includes node id and the different labels. The edge table includes the edge id and the end nodes.

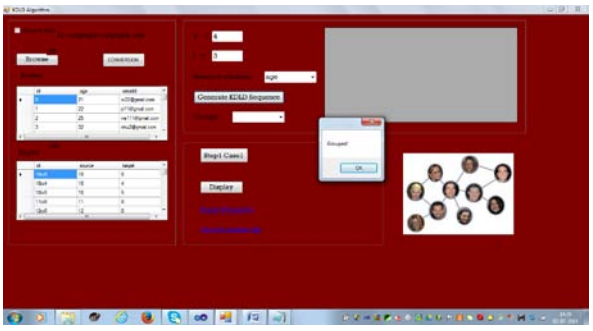

Figure 3.7: Data conversion

3.6.2 Step 2: Grouping

The dataset is divided into no of groups which satisfies three constraints.

1. All the elements in the group have same degree.
2. At least size k
3. The groups have at least l distant label values shows in fig 3.8. After that they perform the APL preserving algorithms. In the APL preserving find the average shortest path length between the nodes in the graph. The five step algorithm is performing in the groups and adding some of the noise node and edges in the graph for generating anonymized dataset.

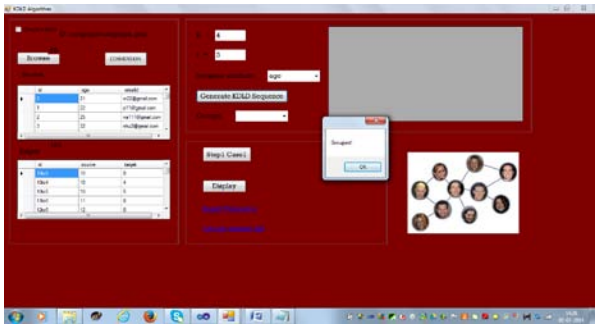

Figure 3.8: Grouping

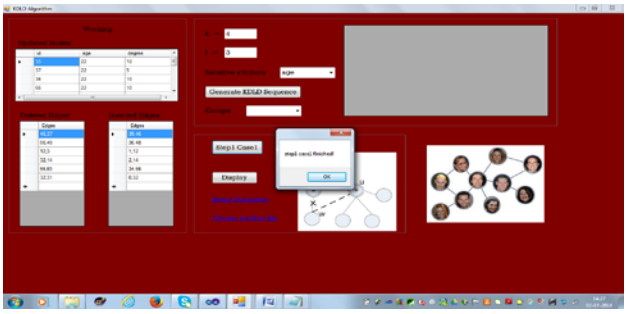

Figure 3.9: Neighbourhood Edge Editing

3.6.3 Step 3 :Adding Node Decrease Degree

In figure 3.10 the Adding node decrease degree step they check whether there exist a node v within two hops of u, and v also needs to decrease its degree. We connect n with v. Since v is within two hops, connecting v with n will not change the distance between u and v.

Figure 3.10: Adding Node Decrease Degree

3.6.4 Step 4: Adding Node Increase Degree

In fig 3.11 the Adding node increase degree step they check whether there exist a node v within two hops of u, and v also needs to increase its degree. We connect n with v. Since v is within two hops of u, connecting v with n will not change the distance between u and v.

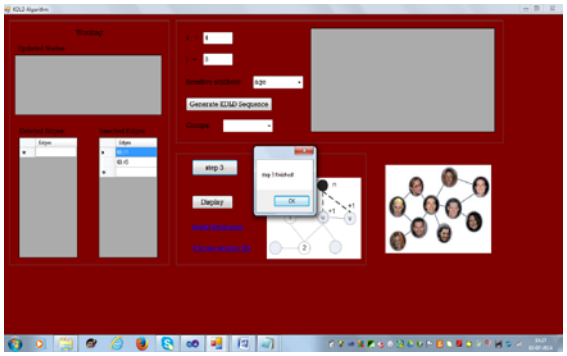

Figure 3.11: Adding Node Increase Degree

3.6.5 Step 5: New Node Degree Setting

In fig 3.12 new node degrees setting them select one edge within the nearest edge set randomly, remove the edge from the graph and connect the endpoints of this edge to the current processing noise node.

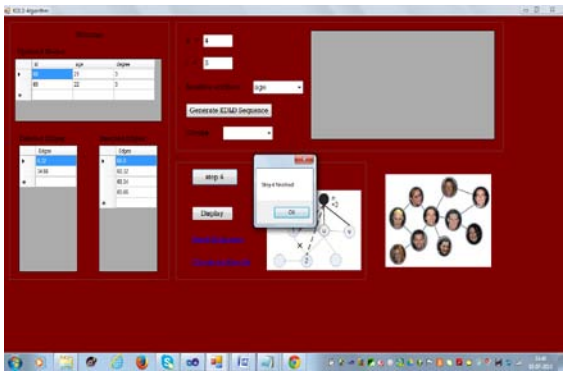

Figure 3.12: New Node Degree Setting

A graph generating software is used for the graph generation. The software takes as input the processed dataset and generates graph. It is based on force directed algorithm and here we used Graph. Graph contains so many algorithm like tree, circle, frequency based etc. So, the graph software is used generate the new graph in different forms. The fig 3.13 shows the graph in FR form.

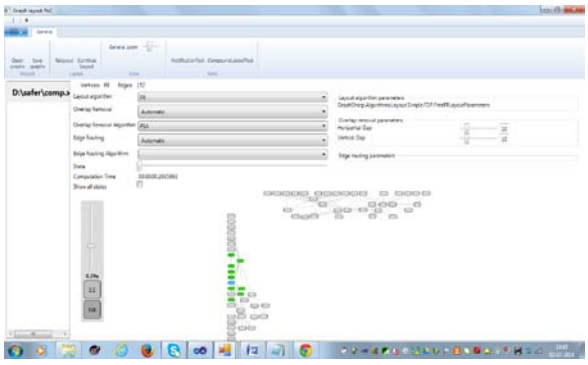

Figure 3.13: Graph in FR Form

4. Conclusion

Privacy is the one of the main challenging issue for publishing the social networking data. The publishing the social networking data they protect the private information of the individuals. In this work, we introduced a new model for the privacy preserving social data in the collaborative environment. The model is called k- degree l - diversity. In this model they protect the private information of individuals from the hackers by adding some noise node in the original graph of individuals and they become modified the graph. First, we have to implement the single source data publishing using this model and then implementing the multisource data publishing. Then we proved that they become high privacy in the multi source data publishing than the single source.

References

[1] L. Sweeney, "K-Anonymity: A Model for Protecting Privacy," Int"l J. Uncertain. Fuzziness Knowledge-Based Systems, vol. 10, pp. 557- 570,2002.
[2] Machanavajjhala, D. Kifer, J. Gehrke, and M.Venkitasubramaniam, "L-Diversity: Privacy Beyond K-Anonymity," ACM Trans. Knowledge Discovery Data, vol. 1, article 3, Mar. 2007.
[3] Vandhana v sree buddha college of engineering kerala university "Anonymization Of Sensitive Labels In Collaborative Data Publishing." I.nternational conference 2014
[4] Zhou and J. Pei, "The K-Anonymity and L-Diversity Approaches for Privacy Preservation in Social Networks against NeighborhoodAttacks," Knowledge and Information Systems, vol. 28, pp. 47-77, 2011.
[5] E. Zheleva and L. Getoor, "Preserving the Privacy of Sensitive Relationships in Graph Data," Proc. First SIGKDD Int"l Workshop Privacy,Security, and Trust in KDD (PinKDD "07), pp. 153-171, 2007.
[6] A. Campan and T.M. Truta, "A Clustering Approach for Data and Structural Anonymity in Social Networks," Proc. Second ACM SIGKDDInt"l Workshop Privacy, Security, and Trust in KDD (PinKDD "08), 2008.
[7] K.B. Frikken and P. Golle, "Private Social Network Analysis: How to Assemble Pieces of a Graph Privately," Proc. Fifth ACM WorkshopPrivacy in Electronic Soc. (WPES "06), pp. 89-98, 2006.
[8] Mingxuan Yuan, Lei Chen, Member, IEEE,PhilipS.Yu, Fellow, IEEE, and Ting Yu" Protecting Sensitive Labels in Social Network Data Anonymization" Ieee Transactions On Knowledge And Data Engineering, VOL. 25, NO. 3, MARCH 2013.
[9] Xintao WuUniversity of North Carolina Xiaowei Ying University of North Carolina at Charlotte Kun LiuYahoo! Labs" A Survey Of Algorithms For Privacy-Preservation Of Graphs And Social Network", 2011.
[10] K.P. Puttaswamy, A. Sala, and B.Y. Zhao, "Starclique: Guaranteeing User Privacy in Social Networks Against Intersection Attacks," ProcFifth Int"l Conf. Emerging Networking Experiments and Technologies (CoNEXT "09), pp. 157-168, 2009