ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 2, Issue 2, pp: (433-438), Month: April-June 2014, Available at: www.researchpublish.com

ANONYMIZATION OF FUNDAMENTAL AND SPREAD SOCIAL NETWORK USING CLUSTERING

G.SAHANA[1], M.MAHESWARI[2], S.SAHANA[3], R.LATHA[4]

Department of Computer Science and Engineering,
Syed Ammal Engineering College, Ramanathapuram.

***Abstract*: Networks structure describes a set of entities and the relations between them. A social network, for example, delivers information on persons in some population and the relations between them, which may describe associations of collaboration, friendship, correspondence and so on. An information network, as extra example, may define technical publications and their reference links. In their most rudimentary form, networks are modeled by a graph, where the nodes of the graph denote to the entities, while edges denote relations between them. Factual social networks may be more composite or contain additional information. Such social networks are of interest to researchers from many disciplines, be it sociology, psychology, epidemiology, or market explore. Nevertheless, the data in such social network cannot be released as such, because it may contain sensitive information. The difficulties of privacy-preservation in social networks rises severely now. We consider the spreaded setting in which the network data is divided between several data holders. The data is divided between a numbers of data holders. The idea is to get there at an anonymized view of the joint network without informative to any of the data holders. A variant of an anonymization algorithm which is based on clustering is used. High sensitive data has been secured in l-diversity algorithm. It provides the learning of confidentiality protection in spread common networks.**

***Keywords*: Social Network, Anonymization, Clustering, Privacy Preserving Data.**

I. INTRODUCTION

An anonymization of the network, is the process of removing identifying attributes like names or social security numbers from the data, is inadequate. The plan behind the attack is to insert a group of nodes with a unique pattern of edges among them into the network. The enemy then may link this unique structure to some set of under attacked individuals. When the anonymized network is released, the enemy traces his injected subgraph in the graph; if successful (that is, there is only one such subgraph in the graph, an occurrence of probability that can be made adequately high), the targets who are linked to this subgraph are re-identified and the edges between them are disclosed. Still less complicated adversaries can use previous knowledge of some possessions of their target nodes (consider, the quantity of their neighbors and their interrelations) in order to recognize them in the released graph and then dig out supplementary information on them.

Therefore, an individual needs to concern a more substantial procedure of anonymization on the network before its release. The methods of privacy protection in networks fall into three main groups. The methodologies of the first group offer *k*-anonymity by means of a deterministic process of edge additions / deletions. In those methods it is considered that the enemy has a background knowledge regarding some property of its target node, and then those methods modify the graph so that it becomes *k*-anonymous in corresponding to that assumed property. The methods of the second group put in noise to the data, by means of switching of edges or, random additions, deletions in order to avoid adversaries from recognizing their target in the network, or inferring the survival of relations between them. The methods of the third group do not alter the graph data like the methods of the two preceding categories; as an alternative, they cluster jointly nodes into super-nodes of range at least *k*, where *k* is the required anonymity perimeter, and then release the graph data in that uncouth resolution.

This work deal with social networks where the nodes can be accompanied by evocative data, and proposes two unique methods of the third group (by clustering the nodes). These algorithms issue anonymized views of the network with extensively smaller information losses than anonymizations issued by the others.

Consider the social network as a plain undirected graph, $G = (V,E)$, where V is the set of nodes and E is the set of edges. Every node corresponds to an individual in the fundamental group, whereas an edge that links two nodes describes a association between the two equivalent individuals. In adding up to the structural data given by E, every node is described by a set of non-recognizing attributes, such as address or zipcode that are known as quasi-identifiers. Combo of such attributes might be used for exclusive identification by linking attacks; hence, they supposed to be widespreaded in order to prevent such attacks. We let $A1, \ldots, An$ indicate the quasi-identifiers, and the set of values that they might attain (e.g., if $A1$ =sexual category then $A1 = \{M, F\}$). After that every node vn, $1 \le n \le N$, is described by a quasi identifier.

II. RELATED WORK

1) Privacy-enhancing k-anonymization of customer data

The k-anonymity form, proposed by Sweeney, is a plain and sensible privacy-preserving model and is expansively studied in recent times. The k-anonymity model assures that every record in the table is indistinguishable to at least (k-1) further records with respect to the privacy-related attributes. Consequently, no privacy related information can be able to infer from the k-anonymity cosseted table during a data mining procedure. It is clear that a analysis classifier can be developed by means of these data to forecast patient's illness based on attributes of Zip, Gender, and Age. If the hospital just releases the table to further organizations for classifier expansion, the organizations may take out patients' disease history by combining this table with other tables.

K-Anonymity Clustering Method

The k-anonymity model, the quasi-identifier characteristic set consists of attributes in a table that potentially describes private information, perhaps by combining with other tables. In adding up, the sensitive attribute is a attribute serves as the class label of all record. The set of three attributes {Zip, Gender, Age} is the quasi-identifier attribute set, while the attribute {Diagnosis} is the sensitive attribute. For every record in the table, its attribute values in the quasi-identifier attribute set are widespread as capsule attribute values, whereas its value of sensitive attribute are not widespread. Throughout generalization, an correspondence class is the set composed of records in the table which has the same values on all attributes in the quasi-identifier attribute set, are assembled to form one correspondence class. The number of records in each correspondence class must be not less than k, which is as the k-anonymity requirement. The value of k is specified by users according to the purpose of their applications. The records in narrative clustering method to create the k-anonymity sheltered table are projected in this work. In the proposed methodology, a prejudiced Attribute C-means clustering algorithm (WF-C-means) is projected to divide all records into correspondence classes. For improving clustering excellence, WF-C-means adaptively adjusts the mass of each quasi-identifier attribute based on the significance of the attribute to clustering quality. The operational procedure in WF-C-means is alike to the C-means algorithm that has good quality scalability for huge data, as a result that the computational competence of WFC- means is feasible in practice. Subsequent to implementing the clustering, a class-merging method merges correspondence classes to make certain that all correspondence classes satisfy the k-anonymity condition. All records in each correspondence class are widespread to be the same with the class middle in the class. Throughout the experiments, the planned clustering method outperforms existing methods in requisites of information deformation measure and computational competence.

2) ℓ-diversity: Privacy beyond k-anonymity

Micro data is a precious source of information for the portion of medical research, public funds, and trend analysis. On the other hand, if folks can be uniquely recognized in the micro data, then their confidential information (such as their medical condition) would be revealed, and this is intolerable. To keep away from the recognition of records in micro data, exclusively recognizing information like social security numbers and names are detached from the table. Though, this first distillation still does not make sure the solitude of individuals in the data.

Bayes-Optimal Privacy

There will likely be numerous adversaries with dissimilar levels of information, each of which is reliable with the full joint allocation. Presume john has a sickness that is (a1) very probable among people in the age group, but (a2) is very uncommon for people of that age group who are teachers. An adversary who simply knows the communication of age and illness will believe that it is very likely for john to have that sickness. On the other hand, an adversary who as well knows that john is a teacher is more likely to consider that john does not have that disease. Thus, though supplementary information can give way improved inferences on standard, there are precise instances wherever it does not. Thus the information publisher has to take into account all probable levels of environmental knowledge.

One the other hand, a number of optimistic disclosures may be satisfactory. For instance, a hospital might be permitted to reveal that a patient has a “kidney problem” since it is well known that the majority patients who visit the hospital have kidney problems. It may also be permitted to reveal that “health Condition” = “well” if this is not considered an attack of privacy. At this end one may be allured to remove tuples with no susceptible “health Condition” values, issue them unaltered, and then generate an miscellaneous version of the residual dataset.

First, releasing unchanged tuples gives an enemy the capability to connect them to outside data and recognize the corresponding folks. This might be measured a privacy break, given that it is sensible for individuals to object to being recognized as correspondents in a review. To keep away from this one might issue a k-anonymous description of tuples with no sensitive “health Condition” values and a different version of the rest of the table. Second, unravelling individuals with no sensitive health conditions from the respite can crash the individuals with responsive health conditions.

3) Preserving the privacy of sensitive relationship in graph data

The objective of data mining is discovering novel and helpful knowledge from data. Every now and then, the data consist of sensitive information, and it wants to be cleaned before it is provided to data mining researchers and the community in order to deal with privacy concerns. Data refinement is a multifaceted problem in which trouncing private information trades off with utility diminution. The aim of refinement is to take away or modify the attributes of the data which help an opponent deduce sensitive information. The answer depends on the properties of the data and the planning of privacy and usefulness in the data.

To formalize privacy protection, Chula et al. Pro-pose a construction based on the instinctive definition that “privacy is confined to the degree we blend in the throng.” What wants to be particularized in this general structure is an consideration of the concept of a database, the opponent in- formation and its functionality, and while an adversary succeeds.

Node anonymization

The anonymization of nodes produces corresponding classes of nodes. Note, though, these corresponding classes are based on node attributes only, and within each correspondence class, there might be nodes with diverse identifying structural characteristics and edges.

Edge anonymization

For instance, we take away the friendship relations in the nodes, because they are the sensitive relations, however we abscond intact the information concerning students taking classes jointly and being scholars of the similar research group. because the relational annotations stay in the graph, this anonymization method should have a high utility. However it is likely to have low privacy protection.

The responsive friendship connection may be re-identified based on node attributes, edge survival or structural characteristics. For instance, consider two scholar nodes containing a Boolean attribute “chatty.” Two nodes that together have it set to “accurate” may be more likely to be friends than two nodes that both have it set to “false.” This inference is based on node attributes. An example of re-identification based on edge existence is two students in the same research group who are more likely to be friends compared to if they are in different research groups. A reidentification that is based on a structural property such as node degree would say that two students are more likely to be friends if they are likely to correspond to high quantity nodes in the graph. A more multifaceted examination is one which uses the consequence of an incidental relationship. For instance, if each of two scholars is highly probably to be a friend with a third person based on other annotations, and then the two scholars are more likely to be friends as well.

4) Graph generation with prescribed attribute constraints

The administration and analysis of social networks has concerned rising interest in the sociology, data mining, database, and theory communities. The majority of previous works are paying attention on revealing attractive properties of networks and discovering proficient and efficient analysis methods. A lot of applications of networks like anonymous Web browsing need relationship anonymity owing to the responsive, stigmatizing, or secret nature of association. It has been exposed in that the plain method of anonymizing graphs by replacing the recognizing information of the nodes with probable ids before releasing the real graph does not assures privacy because the recognition of the vertices can be critically jeopardized by applying sub graph queries. As a consequence, link randomization was recommended. Though, link randomization may considerably influence the utility of the published randomized graph. To conserve utility, we anticipate certain collective characteristics of the unique graph should stay basically unaffected.

Link privacy analysis

We monitor that preserving some attribute in the released graph can considerably breach the privacy, whereas preserving others could not. We should also peak out that, one attribute that jeopardizes solitude in one graph do not essentially jeopardize privacy in another.

We present two graph generation algorithms for the idea of arithmetical testing. In the arithmetical testing, the graph making has stricter necessities. For instance, the producer should be able to access all prospective graphs so that the testing consequence is not prejudiced. In a number of other cases, the attribute principles of the generated graphs are supposed to follow some prearranged distribution. All these troubles involve building a Markov chain with a requisite stationary distribution. The Metropolis-Hastings methodology is one of the standard methodologies of converting a Markov chain with one stationary allocation to another Markov chain with a different stationary allocation.

Switching method applies a Markov chain to produce an artificial graph by switching edges from the novel graph. Switching itself cannot produce consistently sampled directed graphs. A “Go with the winners” algorithm which is based on a non-Markov chain Monte Carlo methodology was proposed to produce uniformly sampled directed graphs. On the other hand, previous switching based algorithms cannot assure the generated graph still conserve some helpful attributes. As shown in this work, a lot of important topological attributes are misplaced in the generated graph. Randomization techniques for testing the importance of discovered patterns have concerned much concentration in data mining [10]. To carry out importance testing of network scrutiny results, it is necessary to generate a group of artificial graphs with attributes fulfilling some distributions. In this work, we offered algorithms based on Markov chain to produce artificial graphs with attribute variety and allocation constraints.

III. EXISTING SYSTEM

The existing system protects the identities of the user under his control from other user. The information that each user needs to protect from other user is the existence or nonexistence. A trusted third party, each user may surrender to him his corresponding partial view of the network. The trusted third party will have a complete view of the entire network, on which he may apply any anonymization algorithm. The Anonymization algorithm will make the sensitive data anonymize in network an give privacy to user data.

The anonymizing of social networks has determined so far is fully on centralized networks, i.e., networks that are seized by one data owner. On the other hand, in a number of settings, the network data is dividing between several data owner or players. For instance, the data in a network of mail accounts where two nodes are associated if the number of email messages that they exchanged was better than some given brink, may be divided between numerous mail service providers. As another instance, assuming a transaction network where an edge represent a monetary transaction between two folks; such a network would be divided between several data holders. In such setups, each individual pedals some of the nodes and he knows only the edges that are adjoining to the nodes under his control.

It is required to work out secure dispersed protocols that would permit the players to pull in at an anonymized account of the combined network. Namely, protocols that would not unveil to any of the interacting individuals more information than that which is disguised by its own effort

IV. PROPOSED SYSTEM

In proposed system, social network user data in stored in separate database. The user could be accompanied by descriptive data, and proposed a novel anonymization method for privacy. Sequential clustering algorithm is used for data clustering which will group the user based on their sensitive data presentation. Anonymization algorithm issues anonymized views of the graph with significantly smaller information losses than anonymizations issued by the other algorithms. L-Diversity algorithm is used to preserve the highly Sensitive data by key generation. The combination L-Diversity algorithm and sequential clustering method give full fledged anonymized view for data holders.

Advantages of Proposed System

By using loss methods information loss can be minimized than other methods that produce the information losses.
In this clustering approach, nodes are clustered based on their neighbourhood values so that efficient cluster will be obtained.
The following diagram shows that how the anonymization can be made.

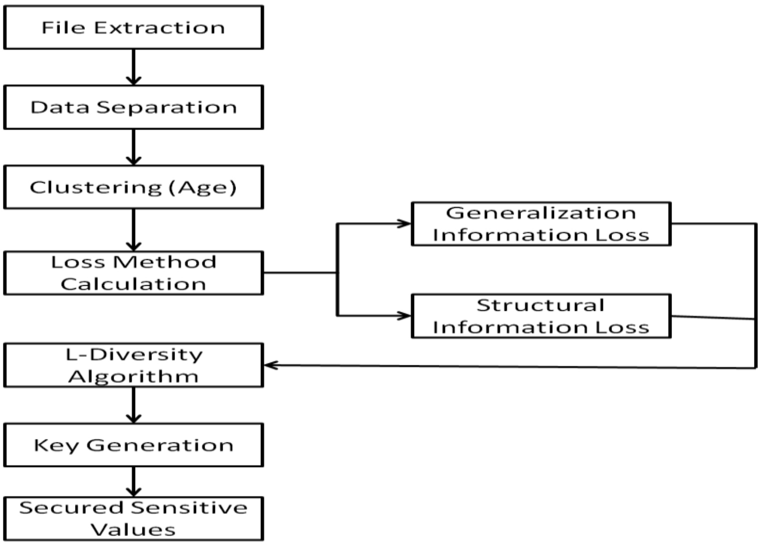

Fig.1 Flow diagram

Data Extraction

The dataset extraction module can be used to extract the dataset and it will be stored in the database for several purposes. At first the dataset was selected, after that it will be split the data separately and it can be stored in multiple tables in the user database.

Clustering Data

In this process, cluster method using here is sequential clustering algorithm. In this clustering algorithm, age has been selected for clustering. After that secure process for male and female candidate has been made. In that secure process, a key is allocated for individual male and female user .This key used for identify male and female users.

Loss Methods Find

In this two types of loss method find

- Generalization Information Loss
- Structural Information Loss

Generalization Information Loss

In this loss method, the loss is measured after the generalization of quasi identifiers such as zipcode and gender.

Structural Information Loss

In this loss method, the loss is measured after the generalization of structure of the network.

Sensitive Data Secure

Sensitive data can be secured by using l-diversity algorithm.
Algorithm:
Input: Social Network.
Output: Social Network with secured data values.

1) Consider the number of attributes to be secured.
2) The attributes will be given as parameter "l" for the diverse process.
3) For each sensitive attribute belong to a user can be produced a unique zipcode.
4) Actual attributes can be replaced using unique zipcode values.

V. CONCLUSION AND FUTURE WORK

To achieve high privacy in social network data L-diversity algorithm is used. It provides security against adversaries by generating keys for sensitive data. Also sequential clustering technique is used to cluster the network data for key generation. Because of finding generalization and structural loss information, the nodes with less loss information can be placed in corresponding cluster. So the enhanced model is produced for preserving privacy of sensitive data in social network. In future work, the security of nodes can be improved by anonymizing in which cluster a node corresponds to.

REFERENCES

[1] Goldberger and T. Tassa. Efficient anonymizations with enhanced utility. *TDP*, 3:149–175, 2010.

[2] X. Wu, X. Ying, K. Liu, and L. Chen. "A survey of privacy-preservation of graphs and social networks". In C. Aggarwal and H. Wang, editors, *Managing and mining graph data*, chapter 14. Springer-Verlag, first edition, 2010.

[3] B. Zhou and J. Pei. Preserving privacy in social networks against neighborhood attacks. In *ICDE*, pages 506–515, 2008.

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

[5] B. Zhou and J. Pei. Preserving privacy in social networks against neighborhood attacks. In *ICDE*, pages 506–515, 2008.

[6] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkita subramaniam." ℓ-diversity: Privacy beyond *k*-anonymity". *ACM Trans. Knowl. Discov. Data*, 1(1):3, 2007.

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

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

[9] M. E. Nergiz and C. Clifton." Thoughts on *k*-anonymization". In *ICDE Workshops*, page 96, 2006.

[10] S. Zhong, Z. Yang, and R. Wright. Privacy-enhancing *k*-anonymization of customer data. In *PODS*, pages 139–147, 2005.