Journal of Theoretical and Applied Information Technology
20th March 2015. Vol.73 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

CLUSTER BASED ANONYMIZATION FOR PRIVACY PRESERVATION IN SOCIAL NETWORK DATA COMMUNITY

[1] V.VIJEYA KAVERI, [2]Dr.V.MAHESWARI

[1]Research Scholar, Sathyabama University , Chennai

[2] Prof., Department of Master of Computer Applications, Sathyabama University , Chennai

E-mail: [1]vijerama.kaveri@gmail.com, [2] mageshwarikarthikeyan@hotmail.com

ABSTRACT

Nowadays, social network data has been made publicly available, analyzed and utilized in one way or another. Since, social network data are publically available it leads to an important issue in Privacy preserving. This paper explores the existing anonymization techniques for privacy preserving publishing of social network data. The problem formulation is done by considering privacy, data utility and knowledge as dimensions in this paper. This paper describes about clustering- based anonymization methods for privacy preservation .

Keywords: *Online Social Networks (OSN) , Privacy Preserving, Anonymization, Clustering.*

1. INTRODUCTION

Online Social Networks(OSNs) are today's popular interactive medium to communicate and share information among different group of users, which provides a way for sharing and consuming useful information. Generally, Social Networks are represented by vertices as actors and edges as relationship between actors. In Our day- to – day lives we come across with mobile networks, friendship networks, academic networks and co-authorship networks.

Social Network Analysis (SNA)[3; 8; 9; 10] has emerged due to swift in social networks. SNA plays an important role in the field sociology, geography, economics, and information science. SNA helps us to analyze the attributes of individual social actors to uncover the hidden social patterns. SNA not only analysis the attributes of individual actors , it also considers the relationship and ties between the social actors in the network.

Online Social Networks is powerful information sharing medium in which privacy is a serious concern when social network data are published. Since the data are available publically a nemesis may intrude privacy of some victims using data and some background knowledge. Mostly the data emerge from settings such as mails, instant messages or mobile calls.

2. CHALLENGES IN ANONYMIZING SOCIAL NETWORK DATA

Even though there are many effective algorithms and models are proposed to preserve privacy in relational data, they are suitable only with relational data alone which is not suitable to social network data.

In order to preserve privacy in published social network data anonymizing is much more challenging than anonymizing relational data [14].

Anonymizing social network data has to face lot of challenges than relational data. In relational data set of attributes are used to associate data from multiple tables where as in a social network graph, subgraphs and neighborhood are used to identify individuals which is much more complicated and much more difficult which leads to model background knowledge of adversaries and attacks about social network data.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Secondly, measuring the information loss is more challenging in social network data than relational data. Generally social network are represented in the form of graph where annonymization is a challenging task when adding or removing vertices and edges which will affect other vertices and edges in a network.

3. MODELING PRIVACY PRESERVATION IN SOCIAL NETWORKS

Generally, models help us to battle privacy attacks and develop protection techniques in social networks. Social Network is modeled as a simple graph with vertices and edges by considering three aspects. First aspect is to identifying the privacy information which may be under attack. Second aspect is based on the knowledge gained, how an hostile may use to attack the privacy of target individuals. Third aspect is how for the data is utilized while the privacy information is fully preserved by applying annonymization method

3.1 Privacy in Social Networks

Normally data's are grouped as sensitive or In-sensitive data, where sensitive data's are considered to be preserved in order to maintain privacy of individuals. In social network data, lot of information is available which are considered as privacy of individuals. Privacy is modeled depending on problem definition, accordingly where different privacy preservation methods are used to model. Some of them are below.

- Vertex existence
- Vertex properties
- Sensitive vertex labels
- Link relationship
- Link weight
- Sensitive edge labels
- Graph metrics.

3.2 Background Knowledge of Adversaries

As social networks are modeled as graph data, the background knowledge of adversaries help us to preserve privacy in publishing social networks. The background knowledge of adversaries may be modeled in various ways.

- Identifying attributes of vertices [1]
- Vertex degrees [4; 5;6;11]
- Link relationship [1;2; 12]
- Neighborhoods[4; 5; 11; 13;14]
- Graph metrics [4; 11]

3.3 Utility in Social Networks

The purpose of anonymzing publishing social network data is to preserve privacy and maximum utilization of data . The anonymized scheme of method vary depending on the type of utilization of data in the application .

Social network data are utilized in general graph properties for example in the distribution of vertex degrees in a network and in aggregate network queries are utilized in customer management applications.

4. CATEGORIES OF ANONYMIZATION METHODS

In order to prevent privacy attacks, the data has to be properly anonymized by selecting a suitable method depending on the model and the utility of the data.

Even though there are challenges in privacy preservation in social network data , anonymization techniques help in developing the privacy models to achieve the goals of anonymized data. Anonymization methods falls into two categories .

4.1 Clustering-based approaches

In this approach social network is represented as graph with vertices and edges which is clustered into groups and anonymizes a subgraph into a super-vertex. In which the details about individuals can be hidden properly. This approach is further classified into vertex clustering methods, edge cluster-ing methods, vertex and edge clustering methods, and vertex-attribute mapping clustering methods.

4.2 Graph modification approaches

In this approach, it anonymizes a graph by modifying edges and vertices in a graph by inserting and/or deleting. This approach is classified into three as Optimization Graph Construction Methods, Randomized Graph Modification Approaches, Greedy Graph Modification Approaches.

5. CLUSTERING-BASED APPROACHES

In this approach the vertices and edges are grouped into clusters and anonymizes a subgraph into a super – vertex . Based on the purpose of clustering it is further classified into four categories as vertex clustering methods, edge clustering methods, vertex and edge clustering methods, and vertex-attribute mapping clustering methods.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

5.1 Vertex Clustering Methods

Vertex clustering approach is identified by Hay et al. [4] in order to address vertex identifier attacks in a simple graph in which vertices and edges are unlabeled. In this method possible background knowledge of an attacker depends on three models of external information, including complete and partial descriptions of vertex neighborhoods, and connections to hubs in the network. Specifically, models are created by using the following queries as the background knowledge of adversaries.

Vertex refinement queries : This class of queries, with increasing attack power, model the local neighborhood structure of a vertex in the network. Consider a vertex v, the weakest knowledge query, denoted as Q0(v), returns the label of v. Since unlabeled graph is considered, Q0(v) returns only in such a case. The queries are successively more descriptive. Q1(v) returns the degree of v, and Q2(v) returns the multiset of degrees of v's neighbors. Generally, query Qi(v) returns the multiset of values which are the results of evaluating Qi-1 on the set of vertices adjacent to v, that is,

$$Qi(v) = \{Qi\text{-}1(u1), Qi\text{-}1(u2), \ldots, Qi\text{-}1(um)\} \quad (1)$$

where u1 …… um are the vertices adjacent to v.

Subgraph queries: This class of queries asserts the existence of a subgraph around the target vertex. The number of edges in the subgraph is used to measure the descriptive power of a subgraph query. An adversary is assumed to be able to gather a fixed number of edges in a subgraph around a target vertex v. By exploring the neighborhood of v, the adversary is capable of identifying whether a subgraph exists around v. The existence of this subgraph can be expressed as a query, and the adversary's knowledge can be modeled by granting the answer to such a query.

Hub fingerprint queries: These classes of queries model the connections of a vertex to a set of selected hubs in the network. A hub is defined as a vertex in a network with high a degree and a high between-ness centrality. Hubs are important components of the topology of networks. A hub fingerprint for a target vertex v is defined as the vector of distances between v and a set of hubs.

5.2 Edge Clustering Methods

Generally, a social network can have multiple types of vertices and edges. Zheleva and Getoor [12] has considered one type of vertices and multiple type of edges in which one edge is considered as sensitive which should be protected against link re-identification attacks. The privacy breach is measured depending on the number of sensitive edges that can be inferred from the anonymized data.

Zheleva and Getoor [12] considered predicting sensitive edges based on the other observed non-sensitive edges in order to model the background knowledge of adversaries. The authors assumed the probabilistic model as the worst case which can predict the existence of a sensitive edge e^{s}_{ij} (that is, an edge between two vertices vi and vj carrying a sensitive label s) based on a set of observations O : $P(e^{s}_{ij}|O)$, where each observation is an edge. In [7] ,a simple noisy-or model is used to identify existence of the sensitive edge. The model can capture the scenario where each observed edge contributes to the probability of the existence of a sensitive edge.

In the nosiy-or model the authors assumed that each observed edge ek has a noise parameter λk,which models the independent influence of ek on the existence of a sensitive edge. The authors assumed that there exists a leak parameter λ0 which models the probability of the existence of a sensitive edge due to some other hidden factors. In this model , the probability of the existence of a sensitive edge is calculated as

$$P(e^{s}_{ij} = 1) = P(e^{s}_{ij} = 1|e1,e2,.., en) = 1 - \prod_{k=0}(1-\lambda k) \quad (2)$$

An adversary succeeds when she/he can correctly figure out whether a sensitive edge exists between two vertices. To achieve more data utility, the authors proposed to count the number of observations which have to be deleted during the anonymization process. The smaller the number of removed observations, the higher the utility. Several graph anonymization strategies are proposed to protect sensitive relationships. In edge anonymization strategy only the sensitive edges are removed all other observed edges remains. In another anonymization strategy only observed edges are removed. Generally, a particular type of observations which significantly contributes to the overall likelihood of a sensitive relationship or a certain percentage of observations that meet some pre-specified criteria (for example, at random,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

connecting high-degree vertices, etc.) can be removed. The most conservative anonymization strategy is to remove all edges in the network. Obviously, in the above approaches, the utility of an anonymized network is low.

In cluster-edge anonymization, the social network is annonymized by collapsing all the equivalence class vertices into a single vertex, and decides which edges to be included in the collapsed graph. A feasible approach is to publish for each edge type the number of edges of the type between two equivalence class vertices.

Cluster-edge anonymization approach aggregates edges on type in order to prevent the disclosure of sensitive relationships. In [4] clusters the vertices to protect vertex identities.

5.3 Vertex and Edge Clustering Methods

Campan and Truta [1] modeled a social network with simple undirected graph where the vertices are associated with some attribute in the network. In the models with relation data, the attributes associated with vertices classified into identifier attributes, quasi- identifier attribute and sensitive attributes as three categories. For example identifier attributes are such as name and SSN which should be removed in publishing, quasi-identifier attributes such as zipcode and sex which may be used by an adversary in reidentification attacks, and sensitive attributes such as diagnosis and income which are assumed to be privacy information. Furthermore, in [1] edges are not labeled.

Campan and Truta [1] , to model data utility they consider the information loss due to generalization and the changes of structural properties. Information loss occurs when vertex labels are generalized. The changes of structural properties quantify the probability of error when one tries to reconstruct the structure of the original social network from the masked version.

Campan and Truta [1] supported the k-anonymity model in order to preserve privacy in social network data. Based on attribute and structural information (neighborhood) every vertex should have at least other (k- 1) vertices. As like relational data, generalization techniques is used to annonymize vertex attribute data and for structural information edge generalization technique is used , which is similar to the one described in [12] to some extent. But in the method, [1] while clustering both the information loss and structural loss are taken into account This process helps to preserve both structural information of the network and vertex attribute information.

5.4 Vertex Attribute Mapping Clustering Methods

Cormode et al. [2] focused on the problem of anonymizing bipartite graphs and he considered the edges as privacy. For example in some applications customers and medical products used are modeled as bipartite graph .

Generally, a bipartite graph G = (U,V,E) consists of |U| vertices of one type and |V| vertices of the other type, and a set of |E| edges $E \subseteq U \times V$. When a bipartite graph is published, the graph structure is retained. The vertices are clustered into groups and the mapping between groups in the original graph and groups in the published graph is released. For example, the mapping table may state that vertices {v1,v2,v3} in the original graph are mapped to {a20, a31, a206} in the published graph. Privacy of entities can be preserved by designing the mapping properly. For example Consumption of the product by the customer can be preserved.

Cormode et al. [2] consider both static attacks and learned link attacks to model the background knowledge of adversaries. If a group of vertices $X \subset U$ only connect to a group of vertices $Y \subset V$, a static attack can immediately obtain the vertices that those in X connect to. Generally, if very few edges exist between vertices in X and vertices not in Y , then a learned link attack can obtain the vertices that those in X connect to with a high confidence.

The accuracy of answering aggregate queries such as the average number of products purchased per user is used for the measurement of data utility. Attributes of vertices in U (or V), or both can be used to compose predicates in aggregate queries, such as the average number of products purchased by customers and the average number of vitamin products by customers.

A safe grouping mechanism is proposed by Cormode et al. [2] protect privacy. A safe grouping of a bipartite graph partitions vertices into groups such that two vertices in the same group of U have no common neighbors in V and vice versa. To control the anonymization granularity, a (k,l)-safe grouping ensures that each group on U contains at least k vertices and each group on V contains at least l vertices.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A safe grouping is identified with the help of greedy algorithm .In which the vertices are processed one by one which may or may not find a safe grouping. The vertices are processed one by one. The algorithm verifies all the vertex and checks whether it can be put into an existing group without breaking the safety. If the vertex is safe then it is added into a group. If the vertex is not safe, a new group is created. After processing all vertices there may be some groups with fewer than m vertices. The algorithm continues with the remaining vertices to run on the collection with a larger group size threshold, say (m+1). Until a safe grouping is found or the group size threshold exceeds the number of vertices in the collection of vertices to be partitioned the iteration continues and finally the algorithm fails in the latter case.

6. CONCLUSION

This paper gives an overview of anonymization techniques for privacy preserving publishing of social network data which is still under research.

This paper gives an idea about how to anonymize social network data to preserve privacy in three dimensions namely, privacy, back ground knowledge, and data utility. Paper tells about the classification of anonymization methods for privacy preservation based on clustering and graph modification approaches.

The paper is finally concluded with the elaborate description of clustering – based annonymization technique which deals with further classification as vertex clustering methods, edge clustering methods, vertex and edge clustering methods, and vertex-attribute mapping clustering methods.

As social network data is more complicated than relational data, it is more challenging and important issue in preserving privacy. Modeling adversarial attacks and developing corresponding privacy preservation strategies are critical.

ACKNOWLEDGEMENTS

I would like to express my gratitude to Sathyabama University for providing me with all the resources necessary for the research.

REFRENCES:

[1] A. Campan and T. M. Truta. A clustering approach for data and structural anonymity in social networks.In Proceedings of the 2nd ACM SIGKDD International Workshop on Privacy, Security, and Trust in KDD (PinKDD'08), in Conjunction with KDD'08, Las Ve-gas, Nevada, USA, 2008.

[2] G.Cormode, D. Srivastava, T. Yu, and Q. Zhang. Anonymizing bipartite graph data using safe groupings. In Proceedings of the 34th International Conference on Very Large Databases (VLDB'08). ACM, 2008.

[3] L. C. Freeman, D. R. White, and A. K. Romney. Research Methods in Social Network Analysis. George Mason University Press, Fairfax, VA, 1989.

[4] M. Hay, G. Miklau, D. Jensen, and D. Towsley. Resisting structural identi¯cation in anonymized social networks. In Proceedings of the 34th International Conference on Very Large Databases (VLDB'08). ACM, 2008.

[5] M. Hay, G. Miklau, D. Jensen, P. Weis, and S. Srivastava. Anonymizing social networks. Technical Report 07-19, University of Massachusetts Amherst, 2007.

[6] K. Liu and E. Terzi. Towards identity anonymization on graphs. In Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data (SIGMOD'08), pages 93{106, New York, NY, USA, 2008.ACM Press.

[7] J. Pearl. Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

[8] J. Scott. Social Network Analysis Handbook. Sage Publications Inc., 2000.

[9] J. Srivastava, M. A. Ahmad, N. Pathak, and D. K.-W.Hsu. Data mining based social network analysis from online behavior. Tutorial at the 8th SIAM International Conference on Data Mining (SDM'08), 2008.

[10] X. Ying and X. Wu. Randomizing social networks: a spectrum preserving approach. In Proceedings of the 2008 SIAM International Conference on Data Mining (SDM'08), pages 739{750. SIAM, 2008.

[11] E. Zheleva and L. Getoor. Preserving the privacy of sensitive relationships in graph data. In Proceedings of the 1st ACM SIGKDD Workshop on Privacy, Security, and Trust in KDD (PinKDD'07), 2007.

[12] B. Zhou and J. Pei. The k-anonymity and l-diversity approaches for privacy preservation in social networks against neighborhood attacks. Submitted for publication.

[13] B. Zhou and J. Pei. Preserving privacy in social networks against neighborhood attacks. In

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Proceedings of the 24th IEEE International Conference on Data Engineering (ICDE'08), pages 506{515, Cancun, Mexico,2008. IEEE Computer Society.

[14] T. M. Galla, Cluster Simulation in Time-Triggered Real-Time Systems, Ph.D. Thesis, Dept. Computer Engineering, University of Technology, Vienna, Austria, 1999.

[15] L. Backstrom, C. Dwork, and J. Kleinberg. Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. In Proceedings of the 16th international conference on World Wide Web (WWW'07), pages 181{190, New York, NY, USA, 2007. ACM Press.