Papri Mani Int. Journal of Engineering Research and Applications www.ijera.com
ISSN : 2248-9622, Vol. 5, Issue 6, (Part -4) June 2015, pp.38-42

RESEARCH ARTICLE OPEN ACCESS

A Brief Survey on Vertex and Label Anonymization Techniques of Online Social Network Data

Papri Mani*, Munmun Bhattacharya**
*(Department of Information Technology, Jadavpur University, Kolkata-98)
** (Department of Information Technology, Jadavpur University, Kolkata-98)

ABSTRACT
With more and more people joining different online social networking (OSN) services every day, the archives of the OSN service providers are increasing drastically. This great amount of personal information is then shared by the service providers with different third parties, which raises a serious concern in preserving privacy of the individuals. For the last few years many work have been done to innovate new techniques, called anonyization techniques, to protect privacy in social network data publishing. In this paper we briefly discuss and categorize vertex and label anonymization techniques which prevent disclosure of individual identities and sensitive information about those identities. We also categorize attributes, attacks and privacy breaches in online social networks.

Keywords – automorphism, equivalence class, isomorphism, k-anonymization, social network graph

I. INTRODUCTION

Online Social Networks are an inseparable part of modern life. They satiate the need and desire of an individual to connect to the rest of the world. A Social Network is represented as a graph where the vertices or nodes represent different real world entities (such as people, organizations or groups) [1] and the edges represent relationships (such as friend, family or colleague) among those entities. People use these Social Networking platforms to connect to their family, friends and colleagues and share their personal views and information with them. So Social Network service providers collect a great amount of private information in their databases. They often share these data with third parties like advertising partners (to get targeted advertisements), application developers and academic researchers. But privacy is a major concern while publishing these data for analysis as an adversary can re-identify a vertex (i.e. an individual), an edge or labels (or attributes) of a vertex using those published data and some background knowledge. In order to stop these privacy breaches many anonymization techniques are adapted while publishing the Social Network data. In this paper we classify the vertex and label anonymization techniques and analyze which kind of privacy attack they prevent.

The rest of the paper is organized as follows: in section II we discuss definitions and notations of a few important terms. In section III and IV we discuss classifications of attributes and privacy breaches in social network data respectively. Section V contains the types of privacy attacks on published social network data. Section VI comprises of categorization of vertex and label anonymization techniques. In section VII we discuss related works. And finally we conclude in section VIII.

II. DEFINITION AND NOTATION

In this section we discuss the definitions and notations of a few terms which are frequently used in rest of the paper.

Definition 1. *Social Network Graph:* a social network graph can be defined as G (V, E, σ, λ), where V is the set of vertices, and each vertex represents an individual in the social network. $E \subseteq V \times V$ is the set of edges (relationships) between vertices, σ is a set of labels that vertices have. $\lambda: V \rightarrow \sigma$ maps vertices to their labels [9]. We use vertex and node interchangeably throughout the paper.

Definition 2. *Equivalence Class:* equivalence class of an anonymized table data is a set of records that have the same values for the quasi-identifiers [6]. But an equivalence class in a social network graph can be defined in terms of quasi-identifiers values or vertex degree or neighbourhood knowledge or a combination of them.

Definition 3. *k-Anonymous Graph:* A graph $G (V, E)$ is said to be k-anonymous if for each vertex $v \in V$ there exist at least other $k-1$ vertices which have either same quasi-identifiers or degree or neighbourhood knowledge as that of v's (i.e. the graph can be divided into a number of equivalence classes having at least k number of vertices each). The process of making a graph k-anonymous is known as k-anonymization.

Definition 4. *Graph Isomorphism:* Let $G = (V, E)$ and $G' = (V', E')$ be two graphs where $|V| = |V'|$. G

and G' are isomorphic if there exists a bijection function f between V and V', $f: V(G) \rightarrow V(G')$, such that $(u,v) \in E$ if and only if $(h(u), h(v)) \in E(G')$. We say that an isomorphism exists from G to G', and $G = G'$. We also say that edge (u, v) is isomorphic to $(h(u), h(v))$. [10]

Definition 5. *k-Isomorphism:* A graph G is k-isomorphic if G consists of k distinct subgraphs $g_1, g_2, ..., g_k$, i.e. $G = \{ g_1, g_2, ..., g_k \}$, where g_i and g_j are isomorphic for $i \neq j$. [10]

Definition 6. *Graph Automorphism:* An automorphism of a graph $G = (V, E)$ is an automorphic function f of the vertex set V, such that for any edge $e = (u, v)$, $f(e) = (f(u), f(v))$ is also an edge in G, i.e., it is a graph isomorphism from G to itself under function f. If there exist k automorphisms in G, it means that there exist k-1 different automorphic functions and G is said to be a ***k*-automorphic** graph.

III. TYPES OF ATTRIBUTES

A social network graph can be labeled or unlabeled. In a labeled social network graph a node has many labels or attributes (like name, age, salary, disease) which represent information about that particular individual. These attributes can be classified into the following three categories [6] with respect to their information revealing capability.

Explicit Identifiers

Some attributes, like *Name*, *Address*, *Social Security Number* etc, can clearly identify an individual. These attributes are known as explicit identifiers.

Quasi Identifiers

These types of attributes when taken alone cannot identify an individual but when taken together they can potentially re-identify a node. For example the attributes *Zip-Code*, *Birth-Date* and *Gender* individually cannot re-identify a node but their values combined together can accomplish the work.

Sensitive Identifiers

Some attributes like *Disease*, *Salary* are considered to be sensitive and confidential. Even attributes from the other two categories can be considered as sensitive with respect to individual preference.

Attributes or labels can be classified into two types with respect to the type of value they hold [6].

Neumerical Attributes

Attributes like *Age*, *Salary*, *Birth-Date* fall under numerical attributes class. In most cases an ordered list is used while generalising these types of attributes.

Categorical Attributes

Attributes like Gender, Country, Work-Class, Education etc fall under this type. Generally a hierarchical list is used while generalising these types of attributes.

IV. TYPES OF PRIVACY BREACHES

There are mainly four types of privacy breaches in online social network data [3] [12].

Vertex Re-Identification

The privacy breach where identity of an individual is revealed is known as vertex re-identification. This re-identification also reveals sensitive labels and all the relationships of the individual under attack with other individuals in the network.

Edge or Relationship Re-Identification

Edge re-identification occurs when relationship between two individuals is revealed. Utilizing social network services (like sending an email or message) generates this kind of information.

Sensitive Label or Attribute Re-Identification

This kind of re-identification leads to revelation of sensitive and confidential attributes, like *Disease*, *Salary* etc., of an individual.

Content Discloser

This kind of breaches disclose the data associated with each vertex, e.g., emails sent and/or received by the individuals in a email network.

V. TYPES OF PRIVACY ATTACKS

Privacy attacks on anonymized social network data can be categorized into three different classes [1] [5]. These kinds of attacks are made by the adversaries with the help of some background knowledge about the target node.

Passive Attack

This kind of attack is attempted after the anonymized social network data is published. Different background knowledge is used in these types of attacks. Neighborhood attack, vertex degree attack, joining attack, structural attack; all fall under this category. Joining attacks are performed by joining two or more anonymized social network databases [7]. For an example a Voter Registration database having attributes name, birth-date, sex and zip-code can be joined with a Hospital Patient database having attributes birth-date, sex, zip-code and disease to re-identify individuals with their diseases (which are sensitive information).

Active Attack

In active attack the adversary embeds a sub-graph i.e. creates new accounts in the social network before the anonymization. The adversary links those new nodes with target nodes. When the anonymized data are published the adversary re-identifies the embedded sub-graph thus re-identifying the target nodes and their position in the social network.

Semi-Passive Attack

No new accounts are created in semi-passive attack but links are created with the target nodes before the anonymization of data.

VI. CATEGORIZATION OF ANONYMIZATION TECHNIQUES

In order to protect the published social network data from the above discussed privacy attacks and prevent leakage of individual personal information, different aspects (vertices, edges and labels) of a social network graph are anonymized before publishing. In this paper different types of vertex and label anonymization techniques are discussed. All of these techniques provide protection against various types of passive attacks.

K-Anonymization

k-anonymization mainly protects against vertex re-identification. It can be achieved by clustering or by modifying the graph and in both the cases generalization and suppression techniques illustrated by Samarati and Sweeney [14] can be used. Though they define the concept on tabular data but it is also used in social network data. In generalization technique the domain of quasi identifier attributes are generalized so that for each node having a general value of an attribute there exist at least *k-1* other nodes with the same general value of the same attribute. Suppression technique is used to remove nodes/tuples from a table before it is published so that a few outliers, i.e. tuples with less than k occurrences, would not force a great amount of generalisation. In [7] LeFevre, DeWitt and Ramakrishnan provide a framework to implement full-domain generalization which is a variation of generalization technique.

There are many clustering methods to achieve *k*-anonymization. Two such methods are enlisted here –

- Vertex clustering: Hay, Miklau, Jensen, Towsley and Weis [11] propose a new method of anonymization which partitions the original graph (an unlabeled graph) by grouping vertices into clusters. Each of the clusters contains at least *k* vertices. Each of the clusters is considered as supernodes and supernodes are connected by superedges which are labeled with non-negative edges. The number of nodes in each cluster and the density of edges that exist within and across the clusters are then published. As only the edge density is published with each supernode, the adversary is unable to distinguish between individuals in a supernode. This anonymization technique generalizes the original graph and protects against vertex re-identification attacks.
- Sequential clustering: In [8] Tassa and Cohen defines a k-anonymization technique using sequential clustering of nodes, which is aimed to protect against link re-identification. They first partition the original network randomly into clusters containing at least k nodes each. The final anonymized graph is formed by checking each node sequentially whether the information loss will be decreased if the node is transferred to another cluster from its current cluster. If the improvement is possible, the node is transferred to the cluster it fits best.

In graph modification technique *k*-aonymization can be achieved in three ways –

- By adding vertices: In [16] Sean Chester et al. propose a method of achieving *k*-anonymization in vertex-unlabeled graphs by adding dummy vertices. They first find out optimal partition for the degree sequence of the graph and then add minimal dummy vertices — so that the distortion to the original graph is minimal — in order to make the output graph *k*-anonymous.
- By adding or deleting edges: In [4] Zhou and Pei use *k*-anonymization to prevent vertex re-identification where the adversary has knowledge about the neighbors of the target vertex and relationship among the neighbors. They consider only *1*-neighbors or immediate neighbors of the target vertex and the subgraph induced by them to be the neighborhood of the target vertex. After extracting neighborhood information of all the vertices they are divided into groups. Neighborhoods of all the vertices in the same group are anonymized by adding/deleting edges so that every vertex has at least *k-1* number of other vertices in the same group with equivalent neighborhood. Graph isomorphism test is used to check equivalence.
- By adding both vertices and edges: Wu et al. [15] insert vertices and edges into a social graph to transform it into a *k*-symmetric graph in order to prevent vertex re-identification. They use automorphism partition to achieve *k*-symmetry.

K-anonymization techniques can also be categorized in terms of structural constraints on vertices like –

- K- Degree anonymization: Liu and Terzi [12] use this method to prevent identity theft in published social network data where the adversary has prior knowledge of degree of target vertex. The graph taken here is undirected, unweighted, containing no self-loops or multiple edges. The anonymization is done in two steps. Firstly the degree sequence *d* of the input graph *G(V,E)* is *k*-degree anonymized to get a new degree sequence d'. And secondly a new graph $G(V, E')$ is constructed with degree sequence d' such that $E' \cap E = E$ (or $E' \cap E \approx E$ in the relaxed version). At first their algorithm uses only edge additions and then the algorithms are

extended to allow simultaneous edge addition and deletion to achieve *k*-degree anonymity.

- K-Isomorphism: Cheng, Fu and Liu [10] use the notion of *k*-security to preserve privacy in social network data. They convert the original graph into a *k*-isomorphic graph as a *k*-secure graph must be *k*-isomorphic. The adversary here has the background knowledge of the subgraph NAG (Neighborhood Attack Graph) which contains the node under attack. This kind of attack is capable of disclosing both vertex and edge information. So to prevent this attack they partition the original graph into *k* subgraphs with same number of vertices and then modify the subgraphs by adding and deleting edges to ensure pairwise subgraph isomorphism.
- K-Automorphism: Zou, Chen and Özsu [17] use *k*-automorphism to protect against structural attacks which lead to vertex re-identification. They convert an original network *G* into *k*-automorphic network *G** where any vertex v in *G** is indistinguishable from its *k-1* symmetric vertices based on any structural information. They use edge addition and edge copy to generate the *k*-automorphic graph.

L-Diversity

k-anonymization provides protection against node re-identification but it is unable to provide sufficient protection against attribute/sensitive label disclosure. Homogeneity attack and background knowledge attack are two attacks which breach the privacy of a *k*-anonymized graph [13]. To address this problem Machanavajjhala, Gehrke, Kifer and Venkitasubramaniam [13] propose a new privacy definition named *l*-diversity. A social network graph satisfies l-diversity when after the partitioning of the graph using k-anonymization each partition contains at least *l* distinct values for the sensitive attribute. In [9] Yuan, Chen, Yu and Yu define *k-degree l-diversity* privacy model to protect against passive attacks where an adversary has degree knowledge of the target vertex. A graph satisfies *k*-degree *l*-diversity if for each vertex in the graph there exist at least *k-1* other vertices with the same degree in the graph and the vertices with same degree contain at least *l* distinct sensitive label values. They use addition/deletion of edges and addition of noise nodes to the original graph to convert it into a *k*-degree *l*-diversity graph. Tassa and Cohen [18] propose *l-sensitive-label-diversity* model to provide personalized sensitive label privacy protection. The adversary here has neighborhood information which includes degree of the target vertex *v* and the labels of *v*'s neighbors. A graph G(V, E) satisfies *l-sensitive-label-diversity* if for each $v \in V$ that associates with a sensitive label, there exist at least *l-1* other nodes with the same neighborhood information, but attached with different sensitive labels.

T-Closeness

Though *l*-diversity combined with *k*-anonymity provides higher level of privacy against sensitive attribute discloser, it is not capable of protecting against all types of attacks. Li, Li and Venkatasubramanian [6] present two such attacks, skewness attack and similarity attack, which can breach *l*-diversity graph. This privacy breach is due to the reason that though *l*-diversity guarantees diversity of sensitive label values in each equivalence class, it does not take into account the semantical closeness of these values. so the authors propose a new technique *t-closeness* which ensures that the distance between the global distribution of a sensitive attribute and equivalence class distribution of the same sensitive attribute is less than or equal to a threshold value *"t"*.

VII. RELATED WORK

Very less work has been done in categorizing privacy preservation techniques in social network data publishing. The survey paper by Zhou, Pei and Luk [2] is the first one to do this type of work, where the authors analyze the privacy models in social networks. They categorize privacy preservation techniques, attacks and background knowledge in social network data. They also give a brief review of the utility of social networks, which is the major concern while anonymizing a social graph. In [5] Soryani and Minaei categorizes the research topics in social network area into seventeen subareas, one of which is "privacy" and their paper focuses on this subarea. They also give a classification of privacy and discuss different aspects, like attack, defence, anonymization etc., related to it. Singh, Bansal and Sofat [3] categorize privacy preservation techniques in social network data publishing by considering two facts – adversary's knowledge and utility of data after release. In [1] Sharma, Mishra, Sharma and Patel classify possible attacks in online social networks into five subcategories, one of which is "traffic analysis attack" and they survey and analyze different techniques to prevent this attack (one of the technique is "Friend in the Middle").

VIII. CONCLUSION

In this paper, we survey different vertex and label anonymization techniques. We also categorize attributes of a node, attacks and privacy breaches in online social networks. Different anonymization technique focuses on protecting different aspects of a social network graph and it is very difficult to find one optimized technique which will cover privacy of all the aspects as well as keep the utility of the published data. Privacy preservation in social network data is more challenging due to the presence of edges in the graph, absence of which makes privacy

preserving in relational data much easier. Because of the graph structure of social network data, an adversary has many different types of information, labels of vertices and edges, degree of nodes, neighborhood graphs and their combinations, to re-identify an individual. So research to improve privacy preservation techniques for online social network data publishing is still in its infancy and needs much more work and exploration.

REFERENCES

[1] M. Sharma, N Mishra, S. Sharma, and R. Patel, "Anonymization in Social Network: Survey and Analysis of various techniques to prevent Traffic Analysis Attack in Online Social Networks," International Conference on Cloud, Big Data and Trust, pp. 202-207, 2013.

[2] B. Zhou, J. Pei, and W. Luk, "A Brief Survey on anonymization Techniques for Privacy Preserving Publishing of Social Network Data," ACM SIGKDD Explorations Newsletter, vol. 10, issue 2, pp. 12-22, December 2008.

[3] A. Singh, D. Bansal, and S. Sofat, "Privacy Preserving Techniques in Social Networks Data Publishing – A Review," *International Journal of Computer Applications (0975-8887), vol. 87- no. 15*, pp. 9-14, February 2014.

[4] B. Zhou and J. Pei, "Preserving Privacy in Social Networks Against Neighborhood Attacks," *Proc. IEEE 24th Int'l Conf. Data Eng. (ICDE '08)*, pp. 506-515, 2008.

[5] M. Soryani and B. Minaei, "Social Networks Research Aspects : A Vast and Fast Survey Focused on the Issue of Privacy in Social Network Sites," *IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 3*, pp. 363-373, November 2011.

[6] N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymization and l-Diversity," IEEE International Conference on Data Engineering, 2007.

[7] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Incognito: Efficient Full-Domain K-Anonymity," ACM SIGMOD International Conference on Management of Data, 2005.

[8] T. Tassa and D. J. Cohen, "Anonymization of Centralized and Distributed Social Networks by Sequencial Clustering," IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 2, pp. 311-324, February 2013.

[9] M. Yuan, L. Chen, P. S. Yu, and T. Yu, "Protecting Sensitive Labels in Social Network Data Anonymization," IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 3, pp. 633-647, March 2013.

[10] J. Cheng, A. W. Fu, and J. Liu, "K-Isomorphism: Privacy Preserving Network Publication against Structural Attacks," SIGMOD, 2010.

[11] M. Hay, G. Miklau, D. Jensen, D. Towsley, and P. Weis, "Resisting Structural Re-identification in Anonymized Social Networks," International Conference on Very Large Data Bases, 2008.

[12] K. Liu and E. Terzi, "Towards Identity Anonymization on Graphs," ACM SIGMOD International Conference on Management of Data, pp. 93-106, 2008.

[13] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-Diversity: Privacy Beyond k-Anonymity," IEEE International Conference on Data engineering, 2006.

[14] P. Samarati and L. Sweeney, "Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression," *Proc. of the IEEE Symposium on Research in Security and Privacy*, 1998.

[15] W. Wu, Y. Xiao, W. Wang, Z. He, and Z. Wang, "K-Symmetry Model for Identity Anonymization in Social Networks," *Proceedings of the 13th International Conference on Extending Database Technology*, pp. 111-122, 2010.

[16] S. Chester, B. Kapron, G. Ramesh, G. Srivastava, A. Thomo, and S. Venkatesh, "k-Anonymization of Social Networks By Vertex Addition," *Proc. 15th ADBIS(2), CEUR Workshop Proceedings*, vol. 789, pp. 107-116, 2011.

[17] L. Zou, L. Chen, and M. T. Özsu, "K-Automorphism: A General Framework for Privacy Preserving Network Publication," International Conference on Very Large Data Bases, 2009.

[18] Y. Song, P. Karras, Q. Xiao, and S. Bressan, "Sensitive Label Privacy Protection on Social Network Data," Scientific and Statistical Database Management, pp. 562-571,2012.