

Giving Privacy to the Sensitive Labels Present on Social Network

Mr.A.K.Baitule, Prof. Manisha Kamble

Nagpur University WCEM, Nagpur, India. Tech (CSE 4th Sem),

Asst. Professor WCEM, Nagpur, India. Nagpur University

akbaitule@gmail.com, manisha.thool@gmail.com

ABSTRACT

This paper aware you why there is need of fine grain & more personalized privacy in data publication on social networks. We propose a privacy protection scheme in which we prevent disclosing identity of users as well as preventing disclosing selected features from users profile. User can select which features of his profile he wants to conceal. We know that on social network data is present in the form of graph in which node represent user & labels represent features. Labels are denoted either as sensitive or as non-sensitive. Node labels are the background knowledge an adversary wants to attack therefore we treat it as protected member. For preventing such type of data we present privacy protection algorithm that allow graph data to be publish in a form such that an opponent who want this information about neighbourhood cannot be safely reveals its identity & its sensitive labels. To this aim, the algorithms transform the original graph into a graph in which nodes are sufficiently indistinguishable by inserting noisy node between them. The algorithms are designed to do so while losing as little information and while preserving as much utility as possible. We evaluate empirically the extent to which the algorithms preserve the original graph's structure and properties.

We show that our solution is effective, efficient and scalable while offering stronger privacy guarantees than those in previous research.

I. INTRODUCTION

There is a chance of privacy threat for publishing data on social network of a particular user so Sensitive information about users of the social networks should be protected. The challenge is to invent methods to publish social network data in such a manner which will not compromising privacy. Various research that was done previously regarding prevention of private information leakage & attacks by malicious adversaries. These early privacy models are mostly concerned with identity and link disclosure. On social network data is present in the form of graph in which node represent user & labels represent feature. The threat definitions and protection mechanisms leverage structural properties of the graph. This paper is motivated by the recognition of the need for a finer grain and more personalized privacy.

Users enter various information like their age, address, current location, occupation or political orientation on social networking sites such as Face book & LinkedIn. These are act as a details & messages in users profile. For that we are proposing protection scheme for private labels which not only prevents disclosing selected features of users profile but also disclosing identity of users. A particular user can select which features or labels of his profile he wish to hide or show.

On social network data is present in the form of graph in which nodes represent user & labels

represent feature. Users can denote each labels as sensitive or insensitive. Figure 1 showing a labelled graph representing a small subset of social network. Each node of the graph indicating users & an edge between nodes represent whether the two users are friend or not. Labels on nodes represent locations of users. Each labels of letter represents city name of each node or users. Some users don't want to disclose their residence but some wants for various reasons. So at the time of releasing these data privacy of their labels should be protected. Therefore the labels indicate as a sensitive or insensitive which are indicated in red colour.

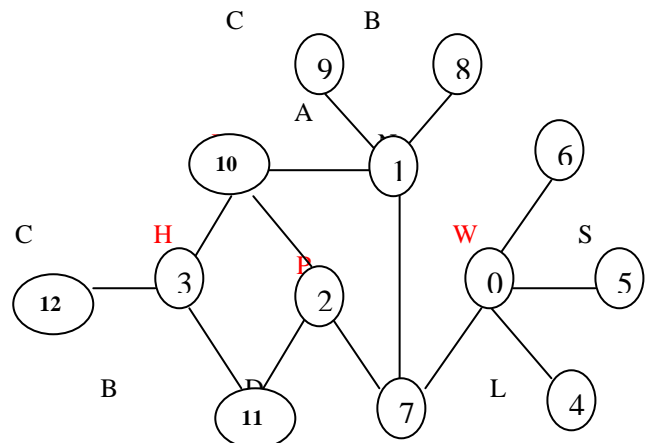


Fig 1. Example of the Labeled Graph representing Social Network

The privacy issue arise when these sensitive labels are disclosing. For not disclosing an alternative is not to enter it on these sites or just delete them. Such a solution reflect an incomplete view of Social Networking websites and it also hide all the information which is linked to that sensitive label which does not disclosing privacy. So an approach is to invent a method prevent sensitive information leakage and also ensuring that identities of particular users are protected from privacy threats. We called these threats as neighbourhood attack, in which an adversary can collect background information like number of neighbours of a node & labels of these neighbours and later on find the sensitive information. For example, if an adversary knows that a user has three friends and that these friends are in A (Alexandria), B (Berlin) and C(Copenhagen), respectively, then she can infer that the user is in H (Helsinki).

The privacy protection mechanism using in our algorithm allow the graph data to be published in such a manner that an adversary cannot access sensitive labels or identity of user. The algorithm transforms the original graph into a graph in which any node with sensitive labels is not distinguishable from other L-1 nodes. The probability to conclude that any node has a sensitive label is not larger than 1/L. For this we design L-diversity-like model, in which we consider node labels as a part of an adversary's background knowledge & as a sensitive information that has to be protected.

The algorithm provides privacy protection by losing as little information & preserving as much utility as possible. In this process of privacy protection algorithm keeps the original graph structure & properties such as density, degree distribution & clustering coefficient. We provide an efficient, effective & scalable solution than the previous research.

The rest of the paper is organized as follows. Section 2 reviews previous works in the area. We define our problem in Section 3 and propose solutions in Section 4. We conclude this work in Section 5.

II. RELATED WORK

The first necessary anonymization technique in both the contexts of micro- and network data consists in removing identification. This nave technique has quickly been recognized as failing to protect privacy. For micro data, Sweeney et al. propose k-anonymity [17] to circumvent possible identity disclosure in naively anonymized micro data. L-diversity is proposed in [13] in order to further prevent attribute disclosure.

Similarly for network data, Backstrom et al., in [2], show that naive anonymization is insufficient as

the structure of the released graph may reveal the identity of the individuals corresponding to the nodes. Hay et al. [9] emphasize this problem and quantify the risk of re-identification by adversaries with external information that is formalized into structural queries (node refinement queries, sub graph knowledge queries). Recognizing the problem, several works [5, 11, 18, 20, 22, 24, 27, 8, 4, 6] propose techniques that can be applied to the naive anonymized graph, further modifying the graph in order to provide certain privacy guarantee. Some works are based on graph models other than simple graph [12, 7, 10, 3].

To our knowledge, Zhou and Pei [25, 26] and Yuan et al. [23] were the first to consider modelling social networks as labelled graphs, similarly to what we consider in this paper. To prevent re-identification attacks by adversaries with immediate neighbourhood structural knowledge, Zhou and Pei [25] propose a method that groups nodes and anonymizes the neighbourhoods of nodes in the same group by generalizing node labels and adding edges. They enforce a k-anonymity privacy constraint on the graph, each node of which is guaranteed to have the same 4 Sensitive Label Privacy Protection on Social Network Data immediate neighbourhood structure with other k-1 nodes. In [26], they improve the privacy guarantee provided by k-anonymity with the idea of l-diversity, to protect labels on nodes as well. Yuan et al. [23] try to be more practical by considering users' different privacy concerns. They divide privacy requirements into three levels, and suggest methods to generalize labels and modify structure corresponding to every privacy demand. Nevertheless, neither Zhou and Pei, nor Yuan et al. consider labels as a part of the background knowledge. However, in case adversaries hold label information, the methods of [25, 26, 23] cannot achieve the same privacy guarantee. Moreover, as with the context of micro data, a graph that satisfies a k-anonymity privacy guarantee may still leak sensitive information regarding its labels [13].

III. PROBLEM DEFINITION

We model a network as $G(V, E, L^s, L, \Gamma)$, where V is a set of nodes, E is a set of edges, L^s is a set of sensitive labels, L is a set of non-sensitive labels and Γ maps nodes to their labels, $\Gamma: V \rightarrow L^s \cup L$. Then we propose a privacy model, L-sensitive-label-diversity; in this model, we treat node labels both as part of an adversary's background knowledge, and as sensitive information that has to be protected. These concepts are clarified by the following definitions:

Definition 1: The neighbourhood information of node v comprises the degree of v and the labels of v's neighbours.

Definition 2. (L-sensitive-label-diversity) For each node v that associates with a sensitive label, there must be at least $L-1$ other nodes with the same neighbourhood information, but attached with different sensitive labels.

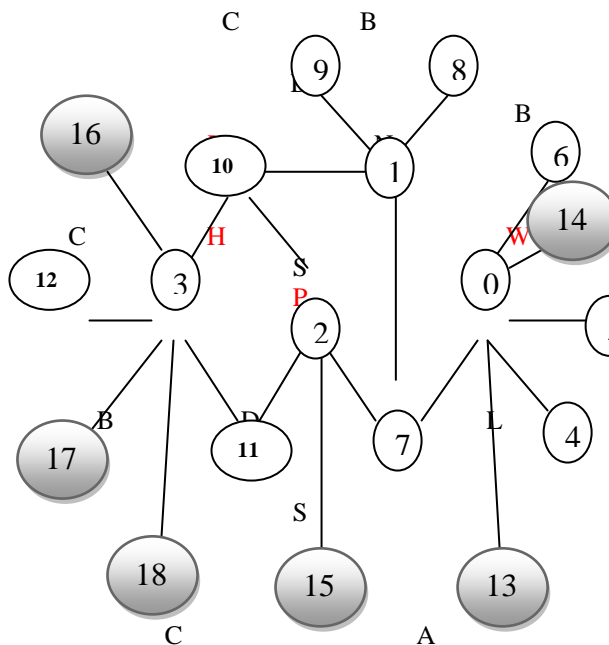


Fig 2 Privacy attaining network example

In Example 1, nodes 0, 1, 2, and 3 have sensitive labels. The neighbourhood information of node 0, includes its degree, which is 4, and the labels on nodes 4, 5, 6, and 7, which are L, S, N, and D, respectively. For node 2, the neighbourhood information includes degree 3 and the labels on nodes 7, 10, and 11, which are D, A, and B. The graph in Figure 2 satisfies 2-sensitive-label-diversity; that is because, in this graph, nodes 0 and 3 are indistinguishable, having six neighbours with label A, B, {C,L}, D, S, N separately; likewise, nodes 1 and 2 are indistinguishable, as they both have four neighbours with labels A, B, C, D separately.

IV. PROPOSED SED WORK

We planned a method which gives the protection to the sensitive labels present on these social networking sites and also if anybody makes an attempt to access these sensitive labels then we will redirect them to the noisy node. When in the network there are sensitive labels present we inserting some noisy node in that network. So the intruder can't get an exact information because noisy node contain false or misleading information.

ALGORITHM

The main objective of the algorithms that we propose is to make suitable grouping of nodes, and appropriate modification of neighbours' labels of nodes of each group to satisfy the L-sensitive-label-diversity requirement. We want to group nodes with as similar neighbourhood information as possible so that we can change as few labels as possible and add as few noisy nodes as possible. We propose an algorithm, Global-similarity-based Indirect Noise Node (GINN), that does not attempt to heuristically prune the similarity computation as the other two algorithms, Direct Noisy Node Algorithm (DNN) and Indirect Noisy Node Algorithm (INN) do. Algorithm DNN and INN, which we devise first, sort nodes by degree and compare neighbourhood information of nodes with similar degree. Details about algorithm DNN and INN please refer to [15].

ALGORITHM GINN

The algorithm starts out with group formation, during which all nodes that have not yet been grouped are taken into consideration, in clustering-like fashion. In the first run, two nodes with the maximum similarity of their neighbourhood labels are grouped together. Their neighbour labels are modified to be the same immediately so that nodes in one group always have the same neighbour labels. For two nodes, v_1 with neighbourhood label set (LS_{v_1}) , and v_2 with neighbourhood label set (LS_{v_2}) , we calculate neighbourhood label similarity (NLS) as follows:

$$NLS(v_1, v_2) = \frac{|LS_{v_1} \cap LS_{v_2}|}{|LS_{v_1} \cup LS_{v_2}|} \quad (1)$$

Larger value indicates larger similarity of the two neighbourhoods.

Then nodes having the maximum similarity with any node in the group are clustered into the group till the group has $\lceil \frac{L}{2} \rceil$ nodes with different sensitive labels. Thereafter, the algorithm proceeds to create the next group. If fewer than $\lceil \frac{L}{2} \rceil$ nodes are left after the last group's formation, these remainder nodes are clustered into existing groups according to the similarities between nodes and groups. After having formed these groups, we need to ensure that each group's members are indistinguishable in terms of neighbourhood information. Thus, neighbourhood labels are modified after every grouping operation, so that labels of nodes can be accordingly updated immediately for the next grouping operation.

This modification process ensures that all nodes in a group have the same neighbourhood information. The objective is achieved by a series of modification operations. To modify graph with as low information loss as possible, we devise three modification operations: label union, edge insertion and noise node addition. Label union and edge

insertion among nearby nodes are preferred to node addition, as they incur less alteration to the overall graph structure.

Edge insertion is to complement for both a missing label and insufficient degree value. A node is linked to an existing nearby (two-hop away) node with that label. Label union adds the missing label values by creating super-values. Sensitive Label Privacy Protection on Social Network Data shared among labels of nodes. The labels of two or more nodes coalesce their values to a single super-label value, being the union of their values. This approach maintains data integrity, in the sense that the true label of node is included among the values of its label super-value. After such edge insertion and label union operations, if there are nodes in a group still having different neighbourhood information, noise nodes with non-sensitive labels are added into the graph so as to render the nodes in group indistinguishable in terms of their neighbours' labels. We consider the unification of two nodes' neighbourhood labels as an example. One node may need a noisy node to be added as its immediate neighbour since it does not have a neighbour with certain label that the other node has; such a label on the other node may not be modifiable, as it is already connected to another sensitive node, which prevents the re-modification on existing modified groups. Sensitive node, which prevents the re-modification on existing modified groups.

Algorithm 1: Global-Similarity-based Indirect Noisy Node Algorithm

Input: graph $G(V, E, L, L^s)$, parameter l ;

Result: Modified Graph G'

```

1 while  $V_{left} > 0$  do
2   if  $|V_{left}| \geq 1$  then
3     compute pairwise node similarities;
4     group  $G \leftarrow v_1, v_2$  with  $Max_{similarity}$ ;
5     Modify neighbors of  $G$ ;
6     while  $|G| < l$  do
7       dissimilarity  $(V_{left}, G)$ ;
8       group  $G \leftarrow v$  with  $Max_{similarity}$ ;
9       Modify neighbors of  $G$  without actually
adding
       noisy nodes ;
10  else if  $|V_{left}| < l$  then
11    for each  $v \in V_{left}$  do
12      similarity  $(v, G_s)$ ;
13       $G_{Max\_similarity} \leftarrow v$ ;
14      Modify neighbors of  $G_{Max\_similarity}$  without
actually
adding noisy nodes;
15 Add expected noisy nodes;
16 Return  $G'(V', E', L')$ ;
```

In this algorithm, noise node addition operation that is expected to make the nodes inside

each group satisfy L-sensitive-label-diversity are recorded, but not performed right away. Only after all the preliminary grouping operations are performed, the algorithm proceeds to process the expected node addition operation at the final step. Then, if two nodes are expected to have the same labels of neighbors and are within two hops (having common neighbors), only one node is added. In other words, we merge some noisy nodes with the same label, thus resulting in fewer noisy nodes.

V CONCLUSION

In this paper we have investigated the protection of private label information in social network data publication. We consider graphs with rich label information, which are categorized to be either sensitive or non-sensitive. We assume that adversaries possess prior knowledge about a node's degree and the labels of its neighbors, and can use that to infer the sensitive labels of targets. We suggested a model for attaining privacy while publishing the data, in which node labels are both part of adversaries' background knowledge and sensitive information that has to be protected. We accompany our model with algorithms that transform a network graph before publication, so as to limit adversaries' confidence about sensitive label data. Our experiments on both real and synthetic data sets confirm the effectiveness, efficiency and scalability of our approach in maintaining critical graph properties while providing a comprehensible privacy guarantee.

REFERENCES

- [1] L. A. Adamic and N. Glance. The political blogosphere and the 2004 U.S. election: divided they blog. In LinkKDD, 2005.
- [2] L. Backstrom, C. Dwork, and J. M. Kleinberg. Wherefore art thou R3579X?: anonymized social networks hidden patterns, and structural steganography. Commun. ACM, 54(12), 2011. Sensitive Label Privacy Protection on Social Network Data 9
- [3] S. Bhagat, G. Cormode, B. Krishnamurthy, and D. S. and. Class-based Graph anonymization for social network data. PVLDB, 2(1), 2009.
- [4] A. Campan and T. M. Truta. A clustering approach for data and structural anonymity in social networks. In PinKDD, 2008.
- [5] J. Cheng, A. W.-C. Fu, and J. Liu. K isomorphism: privacy-preserving network publication against structural attacks. In SIGMOD, 2010.
- [6] G. Cormode, D. Srivastava, T. Yu, and Q. Zhang. Anonymizing bipartite graph data using safe groupings. PVLDB, 19(1), 2010.

- [7] S. Das, Egecioglu, and A. E. Abbadi. Anonymizing weighted social Network graphs. In ICDE, 2010.
- [8] A. G. Francesco Bonchi and T. Tassa. Identity obfuscation in graphs through the information theoretic lens. In ICDE, 2011.
- [9] M. Hay, G. Miklau, D. Jensen, D.Towsley, and P. Weis. Resisting Structural re-identification in anonymized social networks. PVLDB, 1(1), 2008.
- [10] Y. Li and H. Shen. Anonymizing graphs against weight-based attacks. In ICDM Workshops, 2010.
- [11] K. Liu and E. Terzi. Towards identity anonymization on graphs. In SIGMOD,2008.
- [12] L. Liu, J.Wang, J. Liu, and J. Zhang. Privacy preserving in social networks against sensitive edge disclosure. In SIAM International Conference on Data Mining, 2009.
- [13] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkatasubramaniam. - diversity: privacy beyond k-anonymity. In ICDE, 2006.
- [14] MPI. <http://socialnetworks.mpi-sws.org/>.
- [15] Y. Song, P. Karras, Q. Xiao, and S. Bressan. Sensitive label privacy protection on social network data. Technical report TRD3/12, 2012.