Prevent Identity Disclosure in Social Network Data Study

Kun Bai, Ying Liu, Peng Liu
College of Information Sciences and Technology
The Pennsylvania State University
University Park, PA, 16802 USA
{kbai, yliu, pliu}@ist.psu.edu

ABSTRACT

Social networks (P2P network, online communications, and mobile computing) have become the global consumer phenomena in recent years. It its early days, few people foresaw that publishing the consumers' data for merely research purpose will plague users today. The widespread deployment of wireless networking, mobile and embedded devices, sensor networks poses even greater risks to consumers' privacy since adversaries have more power and resources to reveal individual's identity and corresponding sensitive information. In this study, we propose a practical method, named *k-DSA*, to battle such attacks. The experimental results that our method advances existing approaches in anonymizing the network data, and the anonymized data are still usable for the research purpose.

1. INTRODUCTION

Social Networks, such as Peer-to-peer (P2P) computing, online communications, and mobile computing have emerged with increasing popularity in nowadays. This phenomena has gained significant attention from both business and scientific research communities that consider it as a popular model of further utilizing Internet information resources. To better understand this phenomena, e.g., finding a value-added business model, data sets from above services often are published by agencies and organizations for either research or other purposes. Preserving privacy in publishing such microdata becomes crucial because the sensitive information of the individuals may be disclosed. Various approaches have been proposed on relational data to ensure the privacy while providing as much data utility for studies as possible. However, research in preserving privacy on non-relational data has not gained the same progress. In addition, most existing approaches to deal with relational data cannot be easily applied to solve the privacy problem in non-relational data.

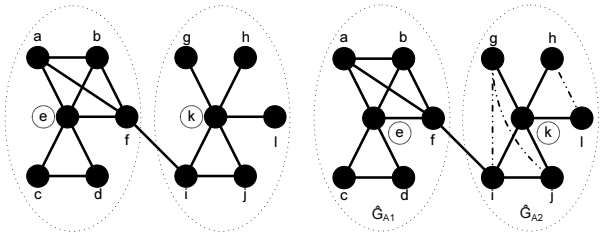
Intuitively, non-relational data (e.g., social networks) are often modeled as complex graphs because graphs are of importance in various applications. Regardless the microdata type, when releasing the microdata, three types of information disclosure have been identified in recent research works: 1) Identity disclosure, in which the identity is linked to a particular individual. 2) Link disclosure (particular in network data and graph model), in which the sensitive relationship (e.g., friendship, patient-doctor relationship) between two individuals are disclosed. 3) Content disclosure, in which the sensitive data (e.g., private blog content, emails, medical records) associated with each individual are leaked. "Apparently, identity disclosure can often cause link disclosure and content disclosure. Once there is identity leak, an individual is re-identified and the corresponding relationship to others and sensitive data are revealed." Thus, intelligent adversaries usually attempt to launch identity disclosure attacks on targeting victims. In this paper, we focus on *identity disclosure* problem in publishing non-relational data.

In practice, external information can be acquired by adversaries in a number of ways, such as adversaries can scan the target network to get prior knowledge about an individual's sensitive attribute, or for social networks, knowledge about existing relationships between known individuals are often publicly available and easily deducible. For networking (e.g., P2P, mobile network) data, an adversary may eavesdrop an individual's communication to recover a list of visited web sites. It is a common vulnerability in network trace collection where an adversary can inject a sequence of identifiable packets. Thus, it is very difficult to bound an adversary's prior knowledge about targeting victims.

2. RELATED WORK

Early works only concern with preserving the privacy of tabular census data. In recent years, privacy increasingly becomes a serious concern in many applications, such as social network, mobile computing and sensor network. Preserving privacy has continuously gained interests in the database and data mining community in studying the complexity of the problem and proposing approaches for anonymizing data records in different anonymization models [5, 3, 1]. However, all these techniques primarily focus on the tabular data. The preserving privacy in non-relational data (e.g., social networks, sensor networks, graphs) has only attract research attention recently.

The work [4] show privacy can still be violated even though the individuals' identities have been hidden. A new privacy protection philosophy called k-anonymity. Many privacy based attacks aim to re-identify individuals via joining the published tabular data with external tables and the background knowledge of individuals. In the k-anonymity scheme, a data set is said to be k-anonymous only if each tuple is indistinguishable from at least other k-1 tuples within the same data set with respect to the *quasi-identifier* attributes. Straightforwardly, better privacy protection can be achieved with a large value of k. However, achieving k-



(a) Original Graph G (b) Anonymized Graph $\widehat{G}$

Figure 1: An Example of K-Dominated-Sets Anonymization. Three edges are added. (The labels for each vertex are just for the sake of illustration and are not in the graphs). (a) has two dominated sets induced to subgraphs $\widehat{H}$={k, g, h, i, j, l}, $\widehat{H}$={e, a, b, c, d, f}, which are isomorphic graphs after the edge operations in (b).

anonymity while satisfying the best data utility is proved to be NP-hard. A number of works [1, 3] have indicated that k-anonymity alone cannot prevent attribute disclosure effectively.

Machanavajjhala indicate in [3] that a k-anonymized dataset has some subtle but severe privacy problems due to the lack of diversity in the sensitive attributes. This observation leads to the proposal of l-diversity principle [3], which advance k-anonymity in protecting against attribute disclosure. However, it also shows several shortcomings as pointed out in [1], for example, it is insufficient to prevent attribute disclosure against similarity attack.

3. PROBLEM DEFINITIONS

In this paper, we model a simple undirected and non-self-looping graph G=(V, E), where V is a set of vertices, $E \subseteq V \times V$ is a set of edges. For a graph G, we use $|V|$, $|E|$ to denote the number of a set of vertices, a set of edges, respectively. V_G, E_G and V, E are interchangeably to denote the set of vertices, the set of edges, respectively. The degree d_{G_v} of a vertex v in a graph G is the number of edges incident to v, also denoted as d_v, equivalently. We use D_G to denote the set of degrees of all vertices in G, d_i is the degree of the i^{th} vertex of G. A degree sequence $\mathbb{D}$ is a set of degrees of the graph G in non-increasing order (e.g., $d_1 \geq d_2 \geq ... \geq d_n$). Two vertices u and v are called adjacent if an edge exists between them. We denote the edge e:{u, v}, or $u \sim v$, and the edge $e \in E$. We now give two important definitions used through this paper.

Definition 1. (*Dominating Set.*) *Given a graph G=(V, E) and an integer $\mathfrak{K} < |V|$, a dominating set $\mathbb{V}'$ for the graph G is a subset $\mathbb{V}' \subseteq V$ such that the following condition is satisfied: $\forall$ vertex $v \notin \mathbb{V}'$, $\exists$ an edge e ={v, v'} $v' \in \mathbb{V}', v \in V$. v' is also called dominating vertex.* □

Definition 2. (*Dominated Set.*) *Given a graph G=(V, E) and its dominating set $\mathbb{V}'$, a dominated set $\underline{\mathbb{V}'}$ of a dominating vertex $v_i^{'} \in \mathbb{V}'$ is a subset $\underline{\mathbb{V}'} \subseteq V$ such that the following condition is satisfied: $\forall$ dominating vertex $v_i^{'} \in \mathbb{V}'$, its dominated set $\underline{\mathbb{V}'}$ contains all vertices $v \notin \mathbb{V}'$ with the radius equal to ϵ to the dominating vertex $v_i^{'}$.*

Definition 3. *(K-Anonymous Graph.) A graph G is k-anonymous if and only if the set of vertex degrees D_G is k-degree anonymous, and the dominated sets $\underline{\mathbb{V}'}$ centered at each dominating vertex $v' \in \mathbb{V}'$ are also k-anonymous for each equivalent class C_i .*□

4. OVERVIEW OF THE APPROACH

In this section, we overview our practical approach to anonymize a network to meet the k-anonymity requirements (Definition 3). This is a multi-steps approach.

First, we form a degree sequence $\mathbb{D}$ from all vertex $v \in G$, then bucket them into m classes each of which contains only distinct degree value. For each class c_i, if the number of vertices $|V_{c_i}| \leqq k$, we construct a new degree sequence locally that is k-anonymous so that each class has at least k vertices and the information loss is minimized.

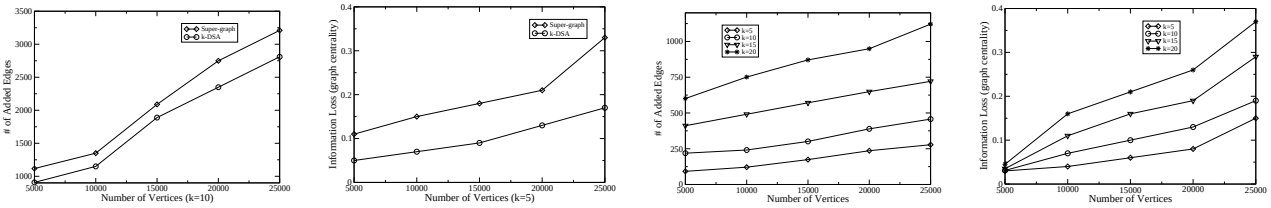
Second, we chose those vertices we consider as dominating vertices and then determine the dominated set of each dominating vertex. These dominated sets cover the entire graph G, and may have overlaps. We than anonymize the dominated sets within each equivalent class starting with those dominating vertices with high degrees because of the common power law of degree distribution, high-degree vertices are always minority.

4.1 Degree Anonymization

Given a degree sequence $\mathbb{D}$ of a graph $G(V, E)$, the algorithm gives a k-anonymous degree sequence $\widehat{\mathbb{D}}$ in which each distinct degree value appears at least k times. We focus on only the edge-addition operations for the sake of simplicity. Hence, we have the anonymized degree sequence $\widehat{\mathbb{D}} = \sum_{i=1}^{|V|} \hat{d}_i \geq \mathbb{D} = \sum_{i=1}^{|V|} d_i$, and for each vertex v_i, $\hat{d}_i \geq d_i$.

Given a sorted (descent) degree sequence $\mathbb{D}$, we divide it into m classes and each of which contains only a unique value. To satisfy the k-degree anonymization that each class should contain at least k times of the unique value, we need to merge together the classes which have a value occurring less than k times to a new class $C_{new} = C_i \cap C_j$, C_i and C_j are adjacent. Recall that one of the goal of k-degree anonymization is to achieve minimal information loss. This can be guaranteed by combining two adjacent classes C_i and C_j, and assigning the vertex degree of all vertices to the highest degree (d_i) in the newly created class C_{new}. The minimal information loss can be met by assigning all degrees in the combined class to the highest degree d_i.

After the combination of non-k-anonymization classes and matching up all vertices in the new class C_{new} to the highest degree, we then have all classes of vertices k-anonymized. However, this k-anonymous degree sequence may not hold the two important properties. Hence, the graph $\widehat{G}$ can not be realized. To realize the degree sequence, we pick up a vertex from a class whose degree value is the lowest among all classes and increase its degree by 1. Consequently, this degree addition makes the previously achieved k-degree anonymization unbalanced. The algorithm then needs to reclassify the unbalanced degree sequence and repeat the entire procedure until both the k-anonymization and the graph $\widehat{G}$ is realizable. We count the increased edges for every vertex and maintain in a hash-table. We do not re-construct an anonymous graph based on the newly generated degree sequence. Instead, we build up the graph at the second step: anonymizing dominated sets.



(a) Super-graph vs. k-DSA, $a_e = 7$

(b) Super-graph vs. k-DSA, $a_e = 7$

(c) k-DSA with Different k, $a_e = 4$

(d) k-DSA with Different k, $a_e = 4$

Figure 2: Evaluation of k-Anonymous Graphs Algorithm

4.2 Dominated Sets Anonymization

We have solved the *k Degree Anonymization* Problem. However, the standalone k-degree anonymization can not protect individual's identity privacy from adversaries by launching structural similarity attacks. Hence, we need to solve the *k-Dominated-Sets Anonymization* problem. In order to meet the k-anonymity requirement, we need to only start with those dominating vertices. Dominating vertices are already classified into each corresponding class. Within each class, our approach makes the dominated sets of each dominating vertex in the same class isomorphic to achieve k-anonymous.

As we have mentioned earlier, finding the dominated sets is relatively straightforward after the dominating set $\mathbb{V}'$ is identified. However, determining whether two subgraphs induced from the corresponding anonymized dominated sets are isomorphic is challenging. In addition, given two simple graphs, determining whether they are isomorphic is NP-hard. In this paper, we use the *adjacency matrix* to present the dominated sets. The adjacency matrix of a finite undirected graph G on n vertices is the n×n matrix where the non-diagonal entry a_{ij} is the number of edges from vertex v_i to vertex v_j. The adjacency matrix of an undirected simple graph is symmetric, and therefore has a complete set of real eigenvalues and an orthogonal eigenvector basis. We first give the definition of isomorphic graphs in our paper as the follows:

PROBLEM 1. *Given two simple graphs G_1 and G_2 with the corresponding adjacency matrices A_1 and A_2, G_1 and G_2 are isomorphic if and only if there exists a* permutation matrix $\mathbb{P}$ *such that*

$$\mathbb{P}A_1\mathbb{P}^{-1} = A_2 \quad (1)$$

then, the problem of proof of isomorphic of two graphs is transformed to finding the permutation matrix of two anonymized dominated sets.

5. EXPERIMENTAL RESULTS

In this section, we demonstrate an empirical study to evaluate our network anonymization approach using both real data sets and synthetic data sets.

5.1 Evaluation of k-Degree Anonymization

We show in Figure 2(a), 2(b) the performance of the k-degree anonymization algorithm on different data sets with different settings of parameter k. In addition, we show in Figure xxx-1 the information loss of anonymization in terms of both the number of edges added and the graph centrality shift. Obviously, when the number of vertices increases, the information loss due to the anonymization increase as well. This is because more edges need to be added to meet the k-degree anonymization requirements. When parameter k increases, the anonymization increases as well because more vertices involve in the k-degree anonymizing. In Figure, we also compare our practical k-degree algorithm with the *Super-Graph* algorithm proposed in [2]. We can observe that our k-degree algorithm outperforms the Super-Graph algorithm with the same parameter settings. Our algorithm guarantees the minimal edges addition and the information loss.

5.2 Evaluation of k-Dominated Sets Anonymization

We show in Figure 2(c), 2(d) the performance of the k-dominated sets anonymization algorithm in terms of the number of added ages, and the information loss (graph centrality skewness, equivalently), where the average vertex degree of the data set is 4. Clearly, we see in Figure 3(e) the number of added edges increase as the number of vertices in the graph G increase, and the number of added edges also increase when the parameter k increase. This is due to the fact that more vertices are involved in the k-anonymity process as the number of vertices and k increase. We see in Figure 3(f) the information loss due to the anonymization. The information loss also increases along with the increasing vertices and the parameter k. Clearly added edges have changed the original graph structure.

6. CONCLUSION

In this paper, we deal with an important problem of preserving privacy in non-relational publication data. We model the non-relational data in a graph and propose a practically feasible solution. An extensive empirical study is conducted on real data sets. The experimental results demonstrate that our solution is highly practical to prevent privacy leakage from the identity disclosure attacks.

7. REFERENCES

[1] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyong k-anonymity and l-diversity. In *Proceedings of the 23rd International Conference on Data Engineering, ICDE 2007*, 2007.

[2] K. Liu and E. Terzi. Towards identity anonymization on graphs. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008*, 2008.

[3] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In *Proceedings of the 22nd International Conference on Data Engineering*, 2006.

[4] L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. In *International Journal on Uncertainty, Fuzziness and Knowledge-based System*, 2002.

[5] L. Sweeney. K-anonymity: a model for protecting privacy. In *International Journal on Uncertainty, Fuzziness and Knowledge-based System*, 2002.