A Privacy Preserving Method for Publishing Set-valued Data and Its Correlative Social Network

Li-e Wang [1] Shan Lin[1] Yan Bai[2] Sang-Yoon Chang[3] Xianxian Li[1] Peng Liu [1*]
[1]Guangxi Key Lab of Multi-source Information Mining & Security, Guangxi Normal University, Guilin, China
[2]School of Engineering and Technology, University of Washington Tacoma, WA 98402, USA
[3]Computer Science Department, University of Colorado Colorado Springs, CO 80918, USA
wanglie@gxnu.edu.cn, lin-sam@foxmail.com, yanb@uw.edu, schang2@uccs.edu, lixx@gxnu.edu.cn, liupeng@gxnu.edu.cn

***Abstract*—Set-valued data and social network provide opportunities to mine useful, yet potentially security-sensitive, information. While there are mechanisms to anonymize data and protect the privacy separately in set-valued data and in social network, the existing approaches in data privacy do not address the privacy issue which emerge when publishing set-valued data and its correlative social network simultaneously. In this paper, we propose a privacy attack model based on linking the set-valued data and the social network topology information and a novel technique to defend against such attack to protect the individual privacy. To improve data utility and the practicality of our scheme, we use local generalization and partial suppression to make set-valued data satisfy the grouped ρ-uncertainty model and to reduce the impact on the community structure of the social network when anonymizing the social network. Experiments on real-life data sets show that our method outperforms the existing mechanisms in data privacy and, more specifically, that it provides greater data utility while having less impact on the community structure of social networks.**

Keywords—Social networks, Set-valued data, Privacy, Data utility, Security

I. INTRODUCTION

Nowadays data mining is becoming increasingly important in many applications. However, many data owners do not have the capability for data mining and rely on the services by third-party institutions, requiring the publishing of the data to the institutions. For example, supermarkets publish the transactional data to data mining companies for customer behavior analysis [1]. Such publication of the original data causes vulnerability against individual privacy [2]. To protect individual privacy of data, a naïve anonymized approach removes the explicit identifying information (e.g., name) before data is released, which by itself is insufficient for individual privacy. Since Fung et al. proposed Privacy-preserving data publishing (PPDP) for protecting data privacy while publishing data [3], PPDP has received significant attention in research communities, and many further protection approaches have been proposed in various data publishing scenarios.

Publishing the set-valued data and its correlative social network data is popular and can collectively provide richer information in many data-mining and social-network applications, such as in behavior prediction and social recommendations [8]-[10]. However, the joint information between the two data sets can provide additional privacy vulnerability which has not been present when considering the two data sets separately. More specifically, the attacker can launch a Linkage Attack (we describe the attack with an illustrative example in Section II). Such attack has not been sufficiently addressed in the previous privacy research in data mining and analyses (we review related work in Section III). Our work is thus motivated by such gap in privacy research.

To resist against the Linkage Attack, we propose a method named Grouped ρ-uncertainty and Anonymized Social Network (Gρ-ASN), which makes set-valued data satisfy grouped ρ-uncertainty and anonymize the correlative social network. To make set-valued data satisfy grouped ρ-uncertainty, we integrate local generalization and partial suppression method in Gρ-ASN. Community structure is an important network property [29]. To make our scheme Gρ-ASN practical, we preserve the community structure while protecting privacy through anonymization, i.e., Gρ-ASN does not need to change the existing community structure.

Our contributions in this work are summarized in the following:

- We propose a new linkage attack model between the set-valued data and its correlative social network when they are published simultaneously, e.g., as illustrated in Example 1 in Section II. We consider the case where both set-valued data and its correlative social network contain sensitive data. The correlation between the data enhances the background knowledge of adversary and yields the vulnerability for the linkage attack. Our attack is distinguishable from the existing studies [7,8,32], which assume that social network data do not contain sensitive information, and only consider them as background knowledge.
- To address the attack model, we develop a novel method called Gρ-ASN which makes the set-valued data satisfy grouped ρ-uncertainty and anonymizes the correlative social network at the same time. For reducing information loss from anonymization leakage, we use local generalization and partial suppression to make set-valued data satisfy grouped ρ-uncertainty.
- For the practical deployment of Gρ-ASN, we preserve the community structue of graph in the process of anonymizing the social network by deleting the intercommunal edges and adding the edges in the communities.

* Corresponding author.

- We validate Gρ-ASN via experiments with two real-world dataset and demonstrate that our approach can effectively reduce information loss and preserve community structure of social network for data utility.

The rest of the paper is organized as follows. Section II describes the Linkage Attack (the motivation of our research) with an illustrative example. Section III presents related work of anonymization of set-valued data and social network. The privacy model is defined in section IV. The algorithm is described in section V. The experimental results are presented in section VI, and finally, section VII concludes this paper.

II. Motivating Example: Linkage Attack

Set-valued data is a common data type and contains sensitive and non-sensitive items. A naïve anonymized transactional data is shown in Table 1(a), in which items *a1*, *a2*, *b1*, *b2*, *b3* are non-sensitive and α, γ are sensitive. Assume that the adversary *Eve* knows the victim *Alice* bought *a2*, so it is easy to infer *Alice* also bought sensitive item α. Resisting against this attack model, ρ-uncertainty model [11][12] protects individual privacy ensuring that the probability of inferring sensitive item set is less than ρ. Table 1 (b) is the anonymized data by means of TDControl from reference [11].

For social networks, graph (consisting of nodes and edges) is popularly used to represent the current state. The nodes represent users and edges represent their friend relationship. A naïve anonymized social network is shown as Fig. 1(a). If the adversary *Eve* knows the *Alice* has three friends, it is easy to re-identify that *Alice* is node 2. To protect privacy of social network, *k*-anonymity [4]-[6]make graph have at least *k* indistinguishable nodes. Fig. 1(b) is the 2-degree anonymous graph using Greedy_Swap from [5].

Table 1. Set-valued Data

(a)Naïve Anonymized Data

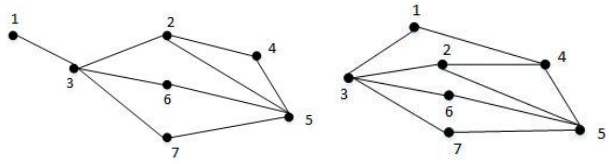
ID	Items
1	a1
2	a2, α
3	b1, b3, α
4	b1, b3
5	a1, b1, α, γ
6	a1, b3
7	a1, b2, γ

(b) TDC(ρ=0.6)

ID	Items
1	a1
2	α
3	b3, α
4	b3
5	a1, α
6	a1, b3
7	a1

However, in order to mine more useful information, set-valued data and its correlative social network data are published simultaneously, such as in behavior prediction and social recommendations [8]-[10]. In this case, the existing single data privacy protection approaches will not be effective as illustrated in our motivating example in Example 1:

Example 1. A Linkage Attack on set-valued and its correlative social network joint publishing. The anonymized transaction set-valued data as Table 1(b) and its correlative anonymous social network as Fig. 1(b) are published simultaneously for data mining. In the set-valued data and its correlative social network, the same ID is corresponding. If an adversary knows *Alice* bought *b3* and had four friends, it is easy to know *Alice* has bought sensitive item α. Although the set-valued data satisfies ρ-uncertainty and the social network is 2-degree anonymous, collectively, they leak information and enables the attacker to track/link between a user and the sensitive item, which would have been difficult if the attacker processed the set-valued data and the social network graph individually.



(a) Naïve Anonymized Graph (b) 2-degree Anonymous Graph

Fig. 1. Correlative Social Network

III. Related Work

Sweeney [12] first proposed the *k*-anonymous model, which makes each record have at least *k* matched records in anonymized dataset. Afterwards, there were numerous related research efforts for individual privacy in set-valued data and social network.

About the privacy of set-valued data, Terrovitis [14] raised the k^m-anonymity model in which the number of records including any *m* items is greater than *k*. To address homogeneity attack, Machanavajjhala [15] proposed *l*-diversity model based on *k*-anonymity, which makes the sensitive itemset have at least *l* different sensitive items in every equivalence group. Sinhong et al. [16] proposed a new method which constructs pseudo taxonomy tree based on utility metrics to anonymize set-valued data. Differential privacy model is used on set-valued data in[17] and[18]. The (*h*, *k*, *p*)-coherence, which is a NP-hard problem proved in [19], ensures that the records including *p* items in anonymized dataset is no less than *k* and the percentage of containing some sensitive items among these record is less then *h*. Furthermore, ρ-uncertainty model was proposed in [11] and modified approaches were proposed in [12][20] by means of partial suppression and local generalization. The ρ-uncertainty protects individual privacy ensuring that the probability of inferring sensitive items is less than ρ. Moreover, Wang et al. [32] considered structural and nonstructural attack and proposed a approximation algorithms based on set-cover greedy approach to achieve *k*-anonymity for set-valued network data. However, they only considered the structural graph data as background knowledge without including sensitive information. Besides, Wang et al. [33] proposed a multifold privacy-preserving model for multi-type data publication. But they didn't consider the jointly publication of set-valued data and social network. In this paper, we consider the set-valued data and its correlative social network both contain sensitive information. We group the records of set-valued data and social network then adopt partial suppression and local generalization to satisfy group anonymity.

In social network, the existing techniques of protecting privacy can be mainly categorized into four types: adding nodes [21][22], adding/deleting edges [23][24], generalization [25][26], and randomization [27]. To avoid node re-identification,[23]and [24] anonymized the social network via adding and deleting edges. In this paper, to prevent the node from getting re-identified in unlabeled graph and preserve the

community structure, we delete these intercommunal edges and add edges in a community first when we anonymize the social network.

In [7] and [8], a privacy-preserving recommendation system was designed to protect the privacy of personalized social recommendation, in which attack can infer the individuals' privacy by observing the victims' rating . However, they concerned about the privacy issues caused by the recommendation algorithm, regardless of the scenario in which the data is published. In this paper, we address the linkage attack between set-valued data and social network while the two kinds of data publishing simultaneously.

IV. Privacy Model and Analyses

A Privacy Concept

As shown in Example 1 in Section II, the linkage attack can associate with the sensitive items of the victim via linking q in D and the degree of node (the victim) in G. Therefore, the individual privacy is compromised when set-valued data and its correlative social network are released. Table 2 gives the definition of variables used in our privacy model.

Table 2. Definition of Symbols

Symbols	Definition
D	The original set-valued data
D′	The anonymized set-valued dataset
q	The non-sensitive items set which the adversary knows
s	A sensitive item
Sup(q)	The number of records that contain q in dataset
Conf(q→s)	The confidence of the rule q→s defined in (1)
G	The original social network
G′	The anonymous social network
V	Nodes set of social network without labels
E	Edges set of social network without labels and weight
C	Communities of original graph
C′	Communities of anonymous graph
T	The set of non-sensitive items in D'
g	The divided groups of D or G

Definition 1 (Sensitive association rule). Given non-sensitive items set q and a sensitive item s in set-valued data, the association rule $q \rightarrow s$ is called a sensitive association rule.

The confidence of rule was defined in [28] for the rule $q \rightarrow s$ defined in the above, the confidence is a probability, which is defined as followed:

$$\mathbf{conf(q \rightarrow s)} = \frac{\mathbf{sup}\ (q \cup s)}{\mathbf{sup}\ (q)} \quad (1)$$

where $\mathrm{sup}(\mathrm{q} \cup \mathrm{s})$ is the number of records including both non-sensitive items q and sensitive item s, and sup(q) is the number of records including q only.

The definition of ρ-uncertainty is as follows: a set-valued data satisfies ρ-uncertainty when the confidence of any sensitive association rules in the released set-valued data is not more than ρ. In the following theorem, we introduce the notion of *grouped ρ-uncertainty*, which is a stronger notion than ρ-uncertainty and requires that all the subsets of the data satisfies ρ-uncertainty.

Theorem 1 (Grouped ρ-uncertainty). Dividing records of set-valued data into groups, if the records in every group satisfy ρ-uncertainty, the total data satisfies grouped ρ-uncertainty.

Proof. Considering set-valued data is divided into groups (g1, g2, … gn), we assume that the confidence of sensitive association rule $q \rightarrow s$ in each group is $(\frac{\mathrm{h1}}{\mathrm{q1}}, \frac{\mathrm{h2}}{\mathrm{q2}}, \ldots \frac{\mathrm{hn}}{\mathrm{qn}})$, which (h1,h2, ... hn) and (q1,q2, … qn) are the number of records in each group simultaneously containing q, s and containing q, respectively. Since the data in every group satisfies ρ-uncertainty, $\frac{\mathrm{h1}}{\mathrm{q1}} < \rho, \frac{\mathrm{h2}}{\mathrm{q2}} < \rho \ldots \frac{\mathrm{hn}}{\mathrm{qn}} < \rho$. Therefore, we have $\mathrm{h1} < \rho \times q1,\ h2 < \rho \times q2 \ldots hn < \rho \times qn$. Adding them together, we get $\mathrm{h1} + \mathrm{h2} \ldots \mathrm{hn} < \rho \times (q1 + q2 + \cdots qn)$. Finally, we get the result of $\frac{\mathrm{h1+h2\ldots\ hn}}{\mathrm{q1+q2+\cdots qn}} < \rho$ proving the theorem. This proof generalizes across the group divisions (g1, g2, … gn) and thus the total set-valued data satisfies ρ-uncertainty.

Definition 2 (Generalization Hierarchy). Generalization Hierarchy is man-made according to the classification of non-sensitive items in set-valued as shown in Fig. 2; for instance, *apple* and *banana* can be generalized to *fruit*.

Definition 3 (Anonymized correlative social network). Assuming that the set-valued data satisfies grouped ρ-uncertainty, nodes of its correlative social network are grouped by the groups of set-valued data. If the nodes in the same group have the same degree, the social network has been anonymized.

Definition 4(Communities in social network). Let G= (V, E) represent a social network, where V is a set of nodes in social network without labels, and $E \in V \times V$ is a set of edges without labels and weight. The nodes in V will be partitioned into communities C ={C1,C2,⋯,Cm}, where $\mathrm{C}_i \cap \mathrm{C}_j = \emptyset$ for all $1 \leq \mathrm{i} \neq \mathrm{j} \leq \mathrm{m}$. For each community in C, the density of internal connections in a community is higher than that of external connections.

In order to preserve the community structure, the community structure of social network was obtained before we anonymize the social network using GN algorithm, a classic community detection algorithm [14]

B Information Loss Metric

To evaluate the effectiveness of anonymous algorithm, we need to analyze the information loss of data. For the set-valued data, we use Normalized Certainty Penalty (NCP) [15] [30] on generalization approach as information loss metric, and the information loss is 1 on suppression approach when an item is suppressed [11][12]. For the social network, we use jaccard similarity[31] which can analyze social network about the community classification accuracy to detect the community similarity between initial graph and anonymized graph. We also adopt the clustering coefficient and average path length to evaluate the utility of graph.

For set-valued data, we use NCP to measure the information loss of item generalization [30]. We define NCP by combining generalization and suppression as follow:

$$\mathbf{NCP(i)} = \begin{cases} \mathbf{1} & \textbf{if i is suppressed} \\ \frac{|\mathbf{m_i}|}{|\mathbf{I}|} & \textbf{if i is generalized to } \mathbf{m_a \epsilon H} \end{cases} \quad (2)$$

where i is a leaf node of m_i in generalization hierarchy H and need to be handled, $|m_i|$ is the number of leaf nodes of m_i, and |I| means the total number of non-sensitive items in D.

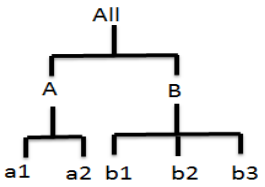


Fig. 2. Item Generalization Hierarchy H

For example, in the item generalization hierarchy in Fig. 2, items $a1$ and $a2$ are generalized to A, while $b1$, $b2$ and $b3$ are generalized to B. If $a1$ is replaced by A and $b1$ is replaced by ALL in the anonymized D', the information loss of items will be computed as NCP($a1$) = $\frac{|A|}{|I|}=\frac{2}{5}$, NCP(b1) = $\frac{|ALL|}{|I|}=\frac{5}{5}=1$. If $a2$ is suppressed, NCP($a2$)=1.

For the total set-valued data, the information loss is defined as follows:

$$\mathbf{NCP}(\boldsymbol{D}) = \frac{\sum_{t \epsilon D} \sum_{x \epsilon t} \boldsymbol{NCP(x)}}{\sum_{t \epsilon D} |\boldsymbol{t}|} \quad (3)$$

here t is a record in D, x is an item in t, and $|t|$ is the total number of items in t.

For social network, the community structure of graph is critical for community analysis. As previously stated, we use jaccard similarity [31] to evaluate the similarity of community between initial graph G and anonymized graph G'. The jaccard similarity is defined as:

$$\mathbf{J}(\boldsymbol{C_i}) = \mathbf{max}\left(\frac{|\boldsymbol{C_i} \cap \boldsymbol{C'_j}|}{|\boldsymbol{C_i} \cup \boldsymbol{C'_j}|} \right), \mathbf{j} \in [\mathbf{1}, \mathbf{m}] \quad (4)$$

where C_i is the community in the original graph G, C={$C1$,$C2$,…,Cn}, C'_j is one community in the communities of anonymized graph as $C' = \{C'_1, C'_2, \ldots C'_m\}$.

To evaluate the similarity of community between G and G', we add the jaccard similarity of each corresponding C_i and C'_j in G and G'.

$$\mathbf{J_{G,G'}} = \frac{\sum_{i=1}^{n} \mathbf{J}(\boldsymbol{C_i})}{\mathbf{n}} \quad (5)$$

V. The Anonymization Algorithm

In this section, we will introduce our novel anonymization algorithm, Grouped ρ-uncertainty and Anonymized Social Network (Gρ-ASN). Gρ-ASN builds on the model and the analyses in Section III. In general, Gρ-ASN consists of three steps. The first step is to divide data into groups and return the ID of data in each group. The second step is to make set-valued data satisfy grouped ρ-uncertainty, which also implies ρ-uncertainty (see Theorem 1). The third step (ASN) is to anonymize the correlative social network, which preserves the community structure of graph well .

A Dividing data into groups

In this step, we need the set-valued data D and the generalization hierarchy H (see Definition 2). Firstly, we find the parents of non-sensitive items in H for each records of D. Then, we put the records whose non-sensitive items have the same parents in H into one group. Finally, we return the ID of records in each groups.

For example, the set-valued data D is showed as table 1(a) and the generalization hierarchy H is showed as Fig. 2. The parents of non-sensitive items in H for each record of D is {{A},{A},{B},{B},{AB},{AB},{AB}}. So the ID in each group [{1,2},{3,4},{5,6,7}] was returned.

B Grouped ρ-uncertainty (GP)

GP: This step makes set-valued data satisfy grouped ρ-uncertainty using Algorithm 1. We need import the set-valued data D, the group g(from Section IV.A), the generalization hierarchy H and the protection strength ρ:

Algorithm 1 GP(D,g,H,ρ)

Input: set-valued data D, group g, generalization hierarchy H and protection strength ρ

1: generalize all non-sensitive items in D to All
2: i=0, n=|g|//n is the group size
3: while i<n do
4: D'←D[g[i]],
5: T←the set of non-sensitive items in D'
6: while T isn't empty do
7: OldNCP=NCP(D')
8: t←randomly select an item from T,
9: T←T-t
10: get generalization rule t→child by H and D
11: D"←update D' by the rule t→child
12: PartialSuppression(D",ρ)
13: NewNCP=NCP(D")
14: if NewNCP<OldNCP do
15: D'=D", T=T+child
16: end if
17: end while
18: update D by D', i++
19: end while

Output: Anonymized D'that satisfies grouped ρ-uncertainty

Lines 1-2 generalize all non-sensitive items in D to All and n is the size of g. Lines 3-17 make each group satisfy ρ-uncertainty. In line *4*, D' is the data in D whose ID in g[i]. Lines 7-17, we generalize every item in T using the top-down generalization method. Line 7 uses NCP(see section IV.B) to compute the information loss of D'. Lines 8-9 randomly selects an item t from T and take t out from T. In line 8, according to H and D, we need get the top-down generalization rule t→child and child is a set, for example if $a1$ and $a2$ are generalized to A, the top-down generalization rule is A→{$a1$,$a2$}. In line 11, child replace t to update D'. Line 12 uses partial suppression method [12] to make D" satisfy ρ-uncertainty. In lines 14-16, if the new NPC of data D" is less

than old NPC of data *D*', we use *D*" replace *D*' and child is added to *T*. In line 18, *D*' replace group *i* data in *D* to update *D*'.

C Anonymized the Correlative Social Network(ASN)

ASN: In Algorithm 2, we anonymize the correlative social network via making nodes in one group have the same degree(Definition 3). In order to reduce the impact on community structure, we aimed to delete the intercommunal edges and add the edge in community.

Algorithm 2 ASN(G,g)

Input: the social network *G*, the group *g*
1: compute average degree of each group in g
2: Avg ← even number of average degree in each group
3: diff← degree of each node in G minus its Avg
4: get communities using GN algorithm,*i*=0,*n*=|*g*|
5: while *i*<*n* do
6: while diff[*i*] > 0 do
7: preserving communities, delete edge E(*i*,*j*)
8: diff [*i*]= diff [*i*]-1 , diff [*j*]= diff [*j*]-1
9: end while
10: *i*++
11:end while
12:while *i*<*n* do
13: while diff[*i*] < 0 do
14: preserving communities, add edge E(*i*,*j*)
15 diff [*i*]= diff [*i*]+1 , diff [*j*]= diff [*j*]+1
16: end while
17: *i*++
18:end while
Output: Anonymized G' that satisfies grouped ρ-uncertainty

In Lines 1-2 of Algorithm 2, we compute average degree of node in every group of g and Avg is the closest even number of average degree of each group. For example, if the average degree of each group is {2, 3, $\frac{8}{3}$ }, the closest even number avg is {2,4,2}. In Line 3, *diff* is a number array that degree of each node in G minus its Avg. Line 4 use GN algorithm [29] to get the communities of graph and *n* is the size of *g*. Lines 5-11 delete edges until every diff of node is not more than 0. In Line 7, in order to preserve communities, we preferentially delete edge E(*i*,*j*) that node *j* and *i* is not in the same community and diff[*j*]>0. Lines 12-18 add edges until all *diff* of node is not less than 0. In line 14, we add edge E(*i*,*j*) which diff[*j*]>0. To preserve communities, node *j* which is in the same community with node *i* is preferential.

D Illustrative Example for Gρ-ASN

We provide an illustrative example of our scheme, Gρ-ASN. Both the set-valued data shown in Table 1(a) and its correlative social network depicted in Fig. 1(a) need to be published for the purpose of studies. To protect the privacy of data, we use Gρ-ASN to protect them. The generalization hierarchy H is built as Fig. 2 and ρ is set to 0.7.

In set-valued data, α and γ are sensitive item while the others are not. Firstly, we need to get the group *g*=[{1,2},{3,4},{5,6,7}] (see setion IV.A). Then using GP, all non-sensitive items are generalized to all as shown in Table 3 (a). In the first layer loop, *D*'=*D*[*g*[0]] and *T*={ALL}. In the second layer loop, OldNCP = NCP(D')= $\frac{2}{3}$, *t*=All, *T*={} and find generalization rule ALL→*A*. Updating *D*' , *D*"={(*A*), (*A*, α)} and *D*" isn't changed using PartialSuppressor method. The variable NewNCP= NCP(*D*")=$\frac{4}{15}$, which is less than OldNCP . So *D*'=*D*"={(*A*), (*A*, α)} and *T*={*A*}. Continually, in the second layer loop, OldNCP=NCP(D')= $\frac{4}{15}$, *t*=*A*, *T*={} and find generalization rule ALL→{*a*1,*a*2}. Updating *D*' , *D*"={(*a*1), (*a*2, α)}, which isn't satisfied ρ-uncertainty. After PartialSuppression method, D" become {(*a*1), (α)} this time. The variable NewNCP= NCP(D")=$\frac{1}{3}$, which is greater than OldNCP, so that D' and T are not changed. For empty T, the second layer loop ends. *D* is updated using *D*' like *g*[0] in Table 3(b). The same steps is used in group *g*[1] and *g*[2] and we get the data like those in Table 3(b).

Table 3. Processing Procedure of Set-valued Data

(a) Grouped ρ-uncertainty

Groups	ID	Items
g[0]	1	ALL
	2	ALL, α
g[1]	3	ALL, α
	4	ALL
g[2]	5	ALL, α, γ
	6	ALL
	7	ALL, γ

(b) PartialSuppression

Groups	ID	Items
g[0]	1	A
	2	A, α
g[1]	3	b1, b3, α
	4	b1, b3
g[2]	5	a1, B, γ
	6	a1, B
	7	a1, B γ

For social networks, we get the groups={{1,2}, {3,4}, {5,6,7}} and get the communities *C*={{1}, {3,6,7}, {2,4,5}} using GN algorithm. The average degree of each group in *g* is {2, 3, $\frac{8}{3}$ } and the closest even number Avg = {2,4,2}. The different value *diff* between degree of each nodes and its *Avg* is [-1,1,0,-2,2,0,0]. We need delete one edge of node 2 and two edges of node 5. We delete edge(2,3) because node 2 and node 3 are not in the same community. For the same reason, we delete edge(5,6) and edge(5,7). After deleting, the *diff* become {-1, 0,-1,-2,0,-1,-1}. For node 1 hasn't other node in its community and edge(1,3) is existing, we have to add edge(1,4). Then we have to add edge(3,4) because edge(3,6) and edge(3,7) are existing. Finally, we add edge(6,7) and the final result is showed as Fig. 3

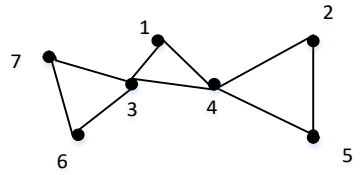


Fig. 3. The Anonymized Correlative Social Network

VI. EXPERIMENTAL Evaluations

Gρ-ASN defends against Linkage Attacks by preventing the attacker to track and link the items in a deterministic manner. To evaluate the effectiveness of our Gρ-ASN algorithm, we experiment on real-world datasets. We further compare the performances of Gρ-ASN with the state of the art algorithms, which conducts anonymization individually on set-valued data (*TDControl [11]* and *Dist [12]*) and social network (*Greedy_Swap* algorithm [5]).

A Datasets and parameters

Our experiments run on two publicly available real-life data sets, crawled online and introduced in [7]: Last.fm and Flixster.

Last.fm, a music service with social network, is a relatively smaller data set. The undirected edges in social network indicate mutual friendship. The directed edges from users to artist indicating listened-to relative. The artists of a user listening to indicate the set-valued record of the user. *Flixster.com*, a movie rating web site with social network, is a relatively larger data set, which contains more than 700k users and approximately 49k movies. Due to the hardware limitations, we chose almost 103k users and all of the movies in our experiments. The rating a user marked for a movie indicates the user has watched the movie so that a set-valued record is the movies the user has watched. The information of the dataset are summarized in table 4.

Our algorithm was implemented with python 2.7 and ran on an Intel Xeon E5-2609 2.10GHz machine with 4GB RAM running Windows 7.

Table 4. Summary of Datasets

Datasets	#|users|	#|edges	#avg.user degree	#|items|
Last.fm	1,892	12,717	13.4	17,632
Flixster	103,271	1,269,076	18.5	48,756

B Data Utility

For set-valued data, we evaluate our algorithm with two aspects: the information loss and the difference number of association rule.

To compute the information loss of data, we use KL(Kullback-Leibler)- divergence [12] to measures the similarity of items distribution between the original data and the anonymized data. The KL-divergence is defined as:

$$\mathbf{KL}(\mathbf{D}, \mathbf{D}') = \sum_i \boldsymbol{D}(\boldsymbol{i}) \log \frac{\boldsymbol{D}(\boldsymbol{i})}{\mathbf{D}'(\boldsymbol{i})} \qquad (6)$$

here D is the original data and D' is anonymized data, D(i) and $D'(i)$ denote the proportion of item i in D and D', respectively.

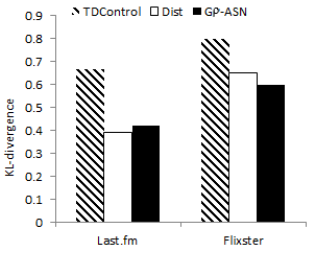
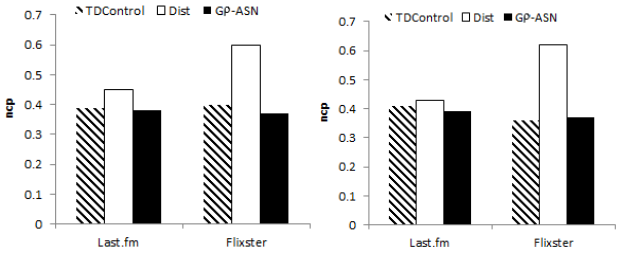


Fig. 4. KL- divergence of Different Dataset

The KL-divergence is showed as Fig. 4. In Last.fm, the information loss of *Gρ-ASN* is much smaller than *TDControl*, but slightly larger than *Dist*. However, *Gρ-ASN* is the best among the three algorithms in the larger data Flixster because we grouped records in the process of privacy protection.

In addition, we use *NCP* (see section IV.B) to measure the information loss of data and the result is depicted in Fig. 5. When ρ is 0.3, the NCP of *Gρ-ASN* is the lowest; when ρ is 0.7, in Flixster, *Gρ-ASN* is a slightly higher than *TDControl*. In general, *Gρ-ASN* has a good performance in the information loss.

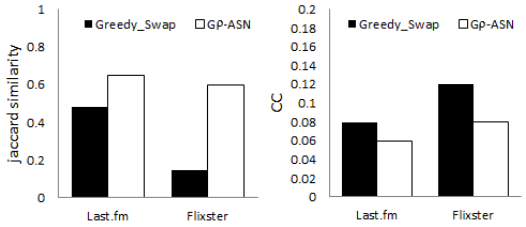


(a) ρ=0.3 (b) ρ=0.7

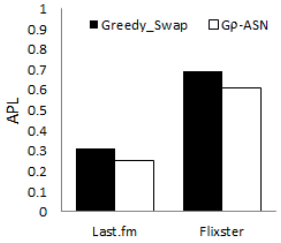
Fig. 5. NCP of Different Dataset

For the social network, comparing with *Greedy_Swap* algorithm(k=15), we use jaccard similarity (see equation 5) to evaluate the community similarity between original graph and anonymized graph. Fig.6(a) shows our algorithm is more similar with original graph since we took some measures on preserving community structure.

Since social networks are complex data, we also use CC (clustering coefficient) and APL (Average Path Length) [5] to evaluate utility of graph. CC represents the degree to which the vertices in a graph tend to be clustered together that exhibits the data feature of community structure. APL measures of the efficiency of information or mass transport on a network. Fig. 6(b) and Fig.6(c) show that our algorithm has a good useful usability.



(a) Jaccard Similarity (b) CC



(c) APL

Fig. 6. Utility of Social Network

VII. CONCLUSION

In this paper, we proposed a linkage attack model where attackers have access to both the set-valued data and its correlative social network and use them collectively to infer and breach individual privacy. To resist such attack, we also proposed a novel approach called Gρ-ASN to anonymize set-valued data and its correlative social network. In contrast to the existing anonymization algorithms, Gρ-ASN defends against the linkage attack. To anonymize the set-valued data, we divide

the records into groups, all of which satisfies ρ-uncertainty. While anonymizing the correlative social network, Gρ-ASN preserves the community structure of social network. The experimental results demonstrate the effectiveness of Gρ-ASN.

For future work, we plan to study the scene of publishing dynamic graph and set-valued data, which can disclose some unexpected data privacy. Due to the dynamic nature of social networks, a more efficient algorithm needs to be developed.

VIII. ACKNOWLEDGMENT

The research was supported by the National Science Foundation of China (Nos. 61662008, 61672176, 61941201 and 61502111), Guangxi "Bagui Scholar" Teams for Innovation and Research Project, the Guangxi Collaborative Center of Multi-source Information Integration and Intelligent Processing, Guangxi Natural Science Foundation (Nos. 2018JJA170082, 2016GXNSFAA380192), and the Guangxi Education Department Project (No. 2018KY0082). This work was also in part supported by the National Science Foundation (NSF) Grant 1921576 and 1922410.

REFERENCES

[1] Liu, J Q. Publishing set-valued data against realistic adversaries. Journal of Computer Science and Technology, 2012, 27(1): 24-36.

[2] Ghinita, G., Tao, Y., Kalnis, P.: On the anonymization of sparse high-dimensional data. In: Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on, IEEE,2008: 715-724

[3] Fung, B.C.M., Wang, K., Chen, R., Yu, P.S.: Privacy-preserving data publishing: A survey on recent developments. ACM Computing Surveys, 2010,42(4).

[4] Zhou B, Pei J, Luk W S. A brief survey on anonymization techniques for privacy preserving publishing of social network data. ACM Sigkdd Explorations Newsletter, 2008, 10(2): 12-22.

[5] Liu, K., Terzi, E.: Towards Identity Anonymization on Graphs. In: ACM SIGMOD International Conference on Management of Data, 2008: 93-106.

[6] Tai C H, Yu P S, Yang D N, et al. Privacy-preserving social network publication against friendship attacks. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. 2011: 1262-1270.

[7] Jorgensen Z, Yu T. A Privacy-Preserving Framework for Personalized, Social Recommendations. In: EDBT. 2014, pp. 571-582.

[8] Zhou P, Zhou Y, Wu D, et al. Differentially private online learning for cloud-based video recommendation with multimedia big data in social networks. IEEE transactions on multimedia, 2016, 18(6): 1217-1229.

[9] Feng X, Sharma A, Srivastava J, et al. Social network regularized sparse linear model for top-n recommendation. Engineering applications of artificial intelligence, 2016, 51: 5-15.

[10] Liu S, Liu A, Liu G, et al. A Secure and Efficient Framework for Privacy Preserving Social Recommendation. Asia-Pacific Web Conference. Springer International Publishing, 2015: 781-792.

[11] Cao J, Karras P, Raïssi C, et al. ρ-uncertainty: inference-proof transaction anonymization. Proceedings of the VLDB Endowment, 2010, 3(1-2): 1033-1044.

[12] Jia X, Pan C, Xu X, et al. ρ-uncertainty anonymization by partial suppression, International Conference on Database Systems for Advanced Applications. Springer, Cham, 2014: 188-202.

[13] Sweeney, L. k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002,10(05): 557-570.

[14] Terrovitis M, Mamoulis N, Kalnis P. Privacy-preserving anonymization of set-valued data. In: Proceedings of the VLDB Endowment, 2008, 1(1): 115-125.

[15] Machanavajjhala, A, Kifer, D, Gehrke, J, , et al. l-diversity: Privacy beyond k-anonymity. ACM Transactions on nowledge Discovery from Data (TKDD), 2007, 1(1): 3-es.

[16] Lin, S, Liao, M. Towards publishing set-valued data with high utility. (2014)

[17] Chen R, Mohammed N, Fung B C M, et al. Publishing set-valued data via differential privacy. Proceedings of the VLDB Endowment, 2011, 4(11): 1087-1098.

[18] Xiao X. Differentially private data release: Improving utility with wavelets and bayesian networks. In: Asia-Pacific Web Conference. Springer, Cham, 2014: 25-35.

[19] Xu Y, Wang K, Fu A W C, et al. Anonymizing transaction databases for publication. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. 2008: 767-775.

[20] Chen, Liuhua, et al. A Sensitivity-adaptive ρ-uncertainty Model for Set-valued Data. International Conference on Financial Cryptography and Data Security. Springer, Berlin, Heidelberg, 2016: 460-473.

[21] Chester, S, Kapron, B, Ramesh, G, Srivastava, G, Thomo, A, Venkatesh, S. k-anonymization of social networks by vertex addition. ADBIS 2(789), 2011, pp:107-116

[22] Jiao J, Liu P, Li X. A personalized privacy preserving method for publishing social network data. In: International Conference on Theory and Applications of Models of Computation. Springer, Cham, 2014: 141-157.

[23] Cheng J, Fu A W, Liu J. K-isomorphism: privacy preserving network publication against structural attacks. In: Proceedings of the 2010 ACM SIGMOD International Conference on Management of data. 2010: 459-470.

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

[25] Campan A, Truta T M. Data and structural k-anonymity in social networks. In: International Workshop on Privacy, Security, and Trust in KDD. Springer, Berlin, Heidelberg, 2008: 33-54.

[26] Wang L, Li X. A clustering-based bipartite graph privacy-preserving approach for sharing high-dimensional data. International Journal of Software Engineering and Knowledge Engineering, 2014, 24(07): 1091-1111.

[27] Boldi, P, Bonchi, F, Gionis, A, Tassa, T. Injecting Uncertainty in Graphs for Identity Obfuscation. Proc. of the VLDB Endowment, 2012,5(11):1376-1387

[28] Verykios V S, Elmagarmid A K, Bertino E, et al. Association rule hiding[J]. IEEE Transactions on knowledge and data engineering, 2004, 16(4): 434-447.

[29] Newman M E J, Girvan M. Finding and evaluating community structure in networks[J]. Physical review E, 2004, 69(2): 026113.

[30] He Y, Naughton J F. Anonymization of set-valued data via top-down, local generalization. Proceedings of the VLDB Endowment, 2009, 2(1): 934-945.

[31] Jaccard P. The distribution of the flora in the alpine zone. 1. New phytologist, 1912, 11(2): 37-50.

[32] Wang S L, Tsai Y C, Kao H Y, et al. Anonymizing set-valued social data. In: 2010 IEEE/ACM Int'l Conference on Green Computing and Communications & Int'l Conference on Cyber, Physical and Social Computing. IEEE, 2010: 809-812.

[33] Wang L, Li X. A graph-based multifold model for anonymizing data with attributes of multiple types[J]. Computers & Security, 2018, 72: 122-135.