A Hybrid Optimization Approach for Anonymizing Transactional Data

Li-e Wang[1,2] and Xianxian Li[1,2](✉)

[1] Guangxi Key Lab of Multi-source Information Mining and Security,
Guangxi Normal University, Guilin 541004, China
[2] College of Computer Science and Information Technology,
Guangxi Normal University, Guilin 541004, China
{wanglie,lixx}@gxnu.edu.cn

Abstract. Transactional data about individuals is increasingly being collected to support many important real-life applications ranging from healthcare to marketing. Thus, privacy issues in sharing transactional data among different parties have attracted considerable research interest in recent years. Due to the high-dimensionality and sparsity of transactional data, existing privacy-preserving techniques will incur excessive information loss. We propose a hybrid optimization approach for anonymizing transactional data through integrating different anonymous techniques. Experimental results verify that our approach significantly outperforms the current state-of-the-art algorithms in terms of data utility.

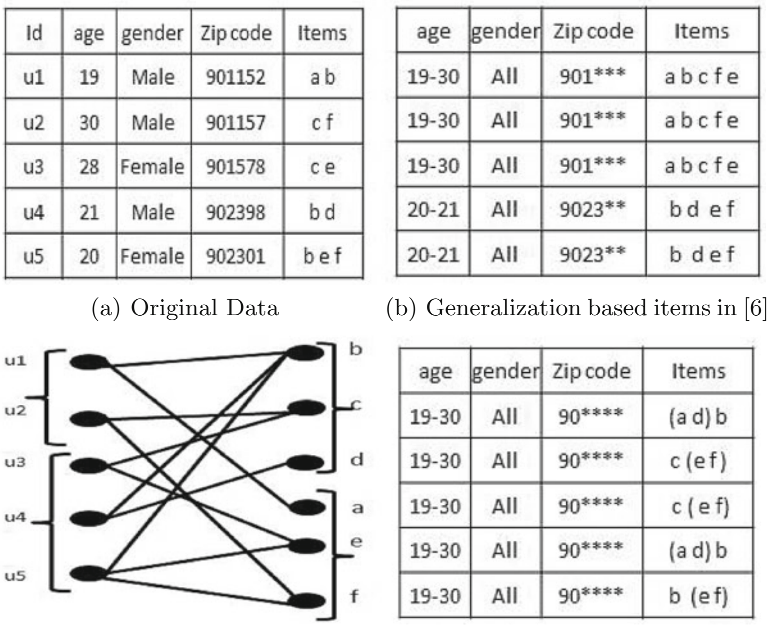
Keywords: Hybrid · Privacy protection · Bipartite graph · Data publishing

1 Introduction

Transactional data, containing information about individuals behaviors or activities, are increasingly used in applications, such as recommendation systems [1], e-commerce [2] and research purposes. Unfortunately, publishing transaction data in its original form may lead to privacy breaches since these data contain individuals private and sensitive information, which contains relational attributes and transaction attributes respectively. Thwarting item disclosure may additionally be needed [3,4]. Due to the high dimensionality and sparsity of transactional data, many methodologies have been proposed to protect the privacy of published transactional data including Generalization which operates by mapping original items to generalized items [4–8] and Suppression which removes items before releasing data [3], Bucketization which operates by separating sensitive items from the QID [9–11] and Perturbation which operates by adding or removing items from individuals transactions [12]. However, these privacy-preserving techniques mostly focus on anonymizing set-valued data only without concerning relational attributes.

G. Wang et al. (Eds.): ICA3PP 2015 Workshops, LNCS 9532, pp. 120–132, 2015.
DOI: 10.1007/978-3-319-27161-3_11

Actually, some applications may require analysis of relational attributes and transaction attributes together. For example, some studies may ask to count all customers above 30 years old who purchased products a and b. Also, some may be interested in analyzing customer demographics and product information together. Purchase records are typical examples of transactional data. Lets take purchase records for example. Purchase records are comprised of transactions, which consist of relational attributes (e.g., attribute information of a customer, such as age, gender and zip code) and transaction attributes (e.g., the purchased products, a set of diseases). For details, see Fig. 1(a). So, a suitable approach to anonymize data having relational attributes and transaction attributes should be needed. Existing works [13,14] proposed multi-dimensional k-anonymization of the whole attributes. There, the problem is that applying those anonymizations, in which they mainly use the approach of generalization to achieve anonymity, to datasets with multiple attributes will lose considerable amount of information.



Id	age	gender	Zip code	Items
u1	19	Male	901152	a b
u2	30	Male	901157	c f
u3	28	Female	901578	c e
u4	21	Male	902398	b d
u5	20	Female	902301	b e f

(a) Original Data

age	gender	Zip code	Items
19-30	All	901***	a b c f e
19-30	All	901***	a b c f e
19-30	All	901***	a b c f e
20-21	All	9023**	b d e f
20-21	All	9023**	b d e f

(b) Generalization based items in [6]

(c) Safe (k,l)-grouping in [15]

age	gender	Zip code	Items
19-30	All	90****	(a d) b
19-30	All	90****	c (e f)
19-30	All	90****	c (e f)
19-30	All	90****	(a d) b
19-30	All	90****	b (e f)

(d) PCTA in [16]

Fig. 1. Examples of anonymization on transactional data

Example 1. *Anonymization on Transactional Database.* Furthermore, we note that different applications (e.g. the personalized recommendation system and the mining of association rules) have different utility requirements. We will give an example to make the point clearer. As of now, most existing recommendation systems use user-based, item-based or collaborative approaches for helping users make decisions. They may require counting sales of products and analyzing customer demographics of products. Yet, another application of association rules are focusing on the relationship among products which are purchased by a customer or one type of customers. That is to say, different applications emphasize different aspects of data, such as statistical characteristics, relationships among products and so on. On the other hand, we notice that these existing anonymized techniques have their own advantages and shortcomings. For detailed explanations, see the Example 1.

As shown in Fig. 1(a), the original database has five transactions and six items. And three different approaches are used to anonymize the same transactional database. Figure 1(b) is 2-anonymous by employing generalization on itemset [6,9], Fig. 1(c) is 2-anonymous by employing a safe (2, 3)-grouping approach [15] and Fig. 1(d) is 2-anonymous by employing Privacy-constrained Clustering-based Transaction Data Anonymization (PCTA) [16]. We can see that Fig. 1(b), (c) and (d) satisfy 2-anonymous according to the model of k-anonymous, but the difference among the results of data utility on the three different published datasets is substantial. For example, a merchandising company wants to figure out what products the customers in particular age ranges prefer to buy. The approach of generalization based items in Fig. 1(b) can provide more information than Safe (k,l)-grouping and PCTA. However, a store wants to watch for the sale of products. The approach of safe (k,l)-grouping based on the bipartite graph in Fig. 1(c) can give an more accurate answer than Generalization and PCTA. The mining of association rules tends to employ the PCTA approach in Fig. 1(d).

Lots of research works has demonstrated that it needs to offer tradeoffs between privacy and utility for applications. Based on above analysis, it is quite clear that, different methods can preserve different aspects of data utility in spite of they all incur a large amount of information loss in terms of transaction attributes or associations among items. Existing methods, such as generalization, clustering, perturbing and suppression, are unable to accommodate specific utility requirements for different applications because they only consider a small number of transformations to anonymize data with multiple attributes. It may cause excessive information loss which would make data useless. Thus, this work proposes a hybrid approach that overcomes the deficiencies of aforementioned anonymized approach to satisfy different utility requirements.

Contributions. To overcome the problems mentioned above, we present a new framework which provides hybrid privacy preserving services based on the form of bipartite graphs via clustering and grouping. The anonymization version of data is presented as a partitioned bipartite graph and an association rules graph. In particular, we focus on different utility requirements of multiple applications. This is crucial difference between our approach and prior works. Note that it is not practical to adopting a single technique for preserving privacy for all different applications since they only consider a small number of transformations to anonymize data. For instance, the method introduced in [15] preserves the degree of nodes perfectly while the method introduced in [16] preserves more information in the mining of association rules. So we propose a hybrid optimization approach to satisfy different utility requirements by integrating different anonymous techniques.

In our framework, we present the anonymized data in a graph form which can handle high dimensional data well. Our approach adopts clustering to anonymize on relational attributes while employing grouping to perturb transaction attributes based on bipartite graphs. Since the approach of grouping items will incur information loss of association rules, we construct a graph of association rules as compensation to satisfy different utility requirements.

We devised an effective anonymization algorithm for generating safe groups which satisfy k-anonymous and l-diversity based on the graph of association rules. We evaluated its performance in real datasets, and experimental results confirm that our approach preserves better data utility to a degree not achieved through previous methods.

Organization. The rest of the paper is organized as follows: Sect. 2 introduces the graph model and utility metrics. Sections 3 and 4 describe our approach and algorithm description in detail. Section 5 demonstrates our approach through experimental study. Section 6 concludes the paper.

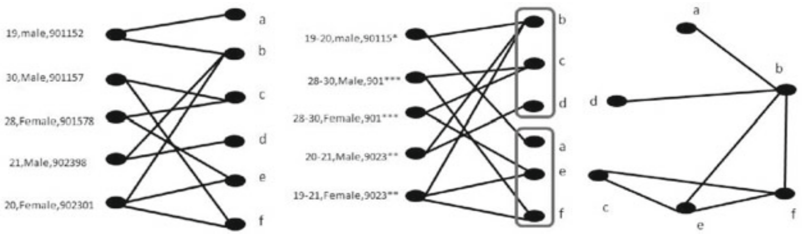
2 Preliminarties

2.1 Graph Model

In this paper, we focus on the problem of anonymizing transactional data and we use bipartite graphs $G = (V, W, E, Lv)$ to simply represent the original data where V and W denote the two types node sets and E denotes the edge set. Each node in the graph has several labels, which represent the QID attributes of the node (such as age, gender and zip code except id). We use Lv to represent the list of labels on nodes. We adopt a graph-based hybrid approach which combines different anonymous techniques to satisfy different utility requirements.

In our framework, we use a partitioned bipartite graph with labels to denote the anonymized graph which achieves anonymity through non-homogeneous generalization [17] and grouping. Meanwhile, we preserve association rules in a weighted graph form. Figure 2 shows a sample instantiation of the schema with Fig. 2(a) showing the original data and Fig. 2(b) showing the correlated bipartite graph with labels and Fig. 2(c) showing the correlated graph of association rules. Here the bipartite graph G in Fig. 2(a) and (b) consists of a set of customer nodes V (such as $u1$,$u2$,$u3$,...) and a set of products nodes W (such as a,b,c,...). An edge $((v \in V, w \in W))$ in E indicates that the customer represented by node v have bought the product represented by node w. In Fig. 2(b), a group is represented by a box and each node in the box belongs to the group.

A graph of association rules is shown in Fig. 2(c). Here the graph G_w (W, E) consists of $n = |W|$ nodes of products and a set of $|E|$ edges. An edge denotes the associations among products. And we use a tuple e= ($w1$, $w2$) to denote an edge from $w1$ to $w2$. If such association incurs more than once, we use a weighted edge



(a) Original Data (b) Partitioned Bipartite Graph (c) Graph of Association Rules

Fig. 2. Examples of hybrid privacy protection

to present the duplicate associations. Each customer can buy different products in various combinations.

2.2 Utility Metrics

The goal of data publishing is to transform the data to generate a publishable version such that (*i*) the *k*-anonymity privacy constraint is satisfied; and (*ii*) utility is maximized. To capture data utility, popular measure include the normalized certainty penalty (NCP) [6,9], which is expressed as the weighted average of the information loss of all generalized items, and the utility loss (*UL*) [16,18], which we use to measure the information loss by generalization on relational attributes in this paper.

Definition 1. Utility loss for non-homogeneous. Given a node $v(l_1,\ldots,l_n)$ and the anonymized node v'$(l_1,\ldots,l_n)$ of v, the node set V and its anonymized set V', the cost of generalization node v to v' and the cost of generalization the whole node set V to V' are measured as follows respectively.

$$UL(v) = \sum_{i=1}^{n} \frac{|l_i'| - |l_i|}{n|V_i|}$$

$$UL(V) = \sum_{v \in V} \frac{UL(v)}{|V|}$$

where $|l_i|$ is the number of distinct values of attribute i in l_i and of attribute i in l_i^* ($l_i^* < l_i$) in V. $|l_i|$ is the number of l_i in V and n is the number of attribute in QID. $|V_i|$ is the number of distinct values of attribute i in V. The utility loss of generalization a node v is the weighted average of the utility loss of generalization all attributes of the node. And the utility loss of generalization the whole node set V is the weighted average of the utility loss of generalization all nodes in the set V. The utility loss of a particular generalization ranges from 0 to 1 and can be easily measured. Obviously, the utility loss of non-homogeneous generalization is less than homogeneous generalization. It says that non-homogeneous generalization can preserve more data utility. Obviously, smaller values of UL indicate less information loss.

On the other hand, information loss of transaction attributes we considering here is incurred by grouping an item to a group. That is, the information loss of transaction attributes is from disturbing a few associations among items. However, the graph of association rules we generated act as a compensation for the information loss which is incurred by grouping. So we omit the information loss of transaction attributes in the experimental section.

Furthermore, we also measure data utility through the quality of answering queries of aggregate analysis as previous work [15]. As defined in [15], we use the parameter of expected error $|\mu - Q|/Q$ to precisely evaluate utility, and the correct answer on original data is Q and the expected answer on anonymized data is μ for each query. Smaller values of the expected error indicate better utility.

3 Achieving Anonymity Through Hybrid Anonymization

As stated above in Example 1, different applications have particular emphasis on different aspects, and different anonymized methods play a different role for preserving data utility. To maximize data utility, we propose a hybrid privacy-preserving approach for transactional data with relational and transaction attributes. The approach adopts non-homogeneous generalization and grouping to preserve the associations between customer nodes and product nodes to guarantee privacy. Meanwhile, we construct a graph of association rules to preserve the associations among products for improving data utility.

3.1 Generating a Graph of Association Rules

This work adopts a methodology of non-homogeneous generalization, which is introduced in [19] and developed in [17], to anonymize relational attributes. Non-homogeneous generalization can improve utility while maintaining an adequate level of privacy. Since non-homogeneous generalization has defined in [17], the details of how to achieve anonymity were omitted for brevity. Note that generalization will lose part of relationships among products, which are bought by a particular individual, since it is privacy. However, the relationships among products are very important for certain applications such as the mining of association rules and may endanger future data collection [20]. For satisfying different utility requirements, we construct a graph of association rules as a part of data publication. The following definition and example illustrates how to generate the graph of association rules.

Definition 2. Graph of association rules. Consider an original database with customer nodes set V and items set W. A graph of association rules is a weighted graphs $G = (W, E)$. Any edge $a \to b \in E$ if and only if, $u \in V$ has bought both a and b. An edge which has more than one occurrence is allowed and represented by weighted edges in graph G.

Example 2. *Constructing a Graph of Association Rules.* Figure 2(c) shows the graph being constructed for Fig. 1(a). The original database contains six items. In the graph, each node denotes an item and each edge denotes the association among items. For example, $u1$ has bought products a and b, which mean that there is an association between items a and b. Thus, they are represented by the edge (a, b) in the graph of association rule. Also like that $u5$ has bought b, e and f, which mean that there are associations among three items b, e and f. Thus, it is represented by three edges (b, e), (b, f) and (e, f), respectively. The assignment routine is called as Assignment edge. Moreover, we allow each association occurring more than once and use weighted edges to represent duplicate edges in the graph. In the process, we recursively invoke Assignment routine on each transaction until all transactions have been processed.

In the graph of association rules, it is easy to find maximum frequent patterns, which is a key problem in data mining research. We generate the graph of association rules as a part of data publication for preserving associations among product nodes, which is lost through grouping and generalization.

3.2 Grouping Based Association Rules

As mentioned above, we guarantee the privacy of relational attributes via generalization and preserve the relationships among products. In this section, we will devise a grouping approach which guarantees transaction attributes based on the association rules. For convenience, we define the relative notions listed below.

Definition 3. *k*-anonymity of attributes. A label of attributes is k-anonymous if and only if there is at least k matched in the entire attribute set for each value.

Definition 4. *k*-anonymity of itemset. A set of items is k-anonymous if and only if the itemset contains at least k individual items.

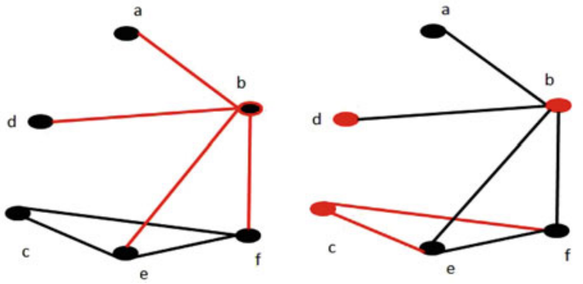
Definition 5. *k*-anonymity in bipartite graph. A bipartite graph is k-anonymous if and only if each label of all nodes attributes is k-anonymous and each itemset is k-anonymous in the graph.

Intuitively, a bipartite graph is k-anonymous if each node is indistinguishable from at least k-1 others. According to above definitions, we know whether a graph is k-anonymous depending on its label and its itemset. The problem of anonymization on relational attributes is addressed by non-homogeneous generalization. This section focus on addressing the privacy problem of transaction attributes by grouping items. For example, there is a 2-anonymous graph (k=2) in Fig. 2(b). The probability of an adversary associating a value of relational attributes, such as aged 20, to an individual in the graph is 1/2. And the probability of an adversary associating an item to a specific node in a group is 1/3. Thus, the bipartite graph satisfies 2-anonymous.

Our approach can be divided into three steps as follows. First, we sort product nodes by its degree in the graph of association rules. Second, we group product nodes based on the graph of association rules and try to find the maximal set of nodes that any edge is non-existent between them. That is, there is no association between the two nodes. So we can prevent the group against homogeneity attack. Let us take a concrete example as an illustration.

Example 3. *Anonymization on items through grouping based on association rules.*

Figure 3 illustrates step by step how the anonymized routine works by grouping nodes into 3-anonymous groups. The original database contains six items, as shown in Fig. 2(c). We prefer to choose the one with the biggest degree as a start point because the biggest degree means that fewer nodes are not associated to it.



(a) Select the start point (b) Reverse Random Walk

Fig. 3. Examples of grouping based on association rules

First, we create a new group A containing b alone, which is considered as a start point on account of its biggest degree 4. Then we walk to all of its neighbors a, d, e, f and find the node c which is not in the neighbor set (see Fig. 3(a)). So c can add into group A. The Anonymize routine is called as reverse random walk. We will check the size of group A after each walk and go on walking until k-anonymous is satisfied. c is considered as a start point in the next walk since it has the second biggest degree in group A. And the next walk is from c to e and from c to f. Then, we can place node a and d into group A when the size of group do not exceed k. Due to the limitation of groups size, we choose d into group A randomly, as shown in Fig. 3(b). In the process, we recursively invoke Anonymize routine on each group by degree until all nodes have been processed. The output of the process is the sequence of groups which satisfy k-anonymous and l-diversity.

Toward different applications such as analyzing frequent pattern of products bought by customers in particular age ranges for recommendation system, counting the sale of products for commercial decisions, or finding maximum frequent patterns for mining association rules, our approach can provide an accurate answer than previous works. That is, our hybrid approach can protect privacy while satisfying different utility requirements.

4 Algorithm Description

In this section, we present a graph-based anonymization algorithm of grouping in our hybrid framework. The algorithm of non-homogeneous generalization was omitted for brevity, which attempts to solve the problem of anonymization on relational attributes. And we will explain the grouping algorithm to tackle the privacy problem of transaction attributes in detail.

Algorithm 1. *Grouping algorithm*

```
Input: $G$, $W$, $k$ //G is the graph of association rule
Output: A $k$-anonymous grouping sequence $GG$.
 1: candidate set $T \leftarrow \emptyset$, temp$\leftarrow \emptyset$, group sequence $GG \leftarrow \emptyset$.
 2: for each $w \in W$ do
 3:    select the node $v$ with the biggest degree in $W$;
 4:    create a new group $g = \{v\}$, $W = W - \{v\}$;
 5:    $T \leftarrow W$;
 6:    for each $w \in T$ do
 7:       if $\exists e(v, w) \in G$ then
 8:          $temp = temp \cup \{w\}$, $T = T - \{w\}$;
 9:       end if
10:    end for
11:    Sort candidate set $T$ by the degree in descending order;
12:    for each $w \in T$ do
13:       if $(|g|_{size} < k)$ then
14:          $g = g \cup \{w\}$, $T = T - \{w\}$, $W = W - \{w\}$;
15:          $n$= next node in $T$;
16:          while $\exists e(w, n) \in G$ do
17:             $n$= next node in $T$;
18:          end while
19:          $w$=$n$;
20:       else
21:          $GG = GG \cup g$;
22:          break;
23:       end if
24:    end for
25: end for
```

The algorithm of how to generate a graph of association rules through grouping is shown in Algorithm 1. The inputs of the algorithm are the original product nodes list of W, its graph of association rules G and the anonymous parameter k. The loop of lines 2–25 tries to find these product nodes that have no association. And the algorithm divides them into a k-anonymous group for resisting against homogeneity attacks. Line 3 selects the node having the biggest degree as a start point. Line 4 creates a new group g containing v alone, and deletes v from W. The candidate set T is set to the universal set W in line 5. The loop of lines 6–10 tries to construct the candidate set which has no association with v. Line 11 sorts all nodes in candidate set T by the degree in descending order. And we continue with sorted node sequence T which can group nodes more effectively. The loop of lines 12–24 tries to generate a k-anonymous group for v from its candidate set T. In detail, line 13 checks whether the size of group g is more than k or not. If the size of group g is less than k, the algorithm will select a node w having the second biggest degree from its candidate set T and add w into group g and delete w from T in line 14. Then the algorithm tries the next node in T. If the size of group g is larger than k, add g into the partition sequence GG. And the process will create a new group and enter the next turn until all nodes in W have been processed. In practice, a greedy algorithm is adopted to generate safe groups in our approach. Although this algorithm is possible to fail, it is easy to generate safe groupings and guarantee privacy on transaction attributes, since it is high-dimensional and sparse.

Differing from previous works, our hybrid framework integrating different anonymous techniques try to satisfy different utility requirements. In detail, we employ partitioning algorithm to achieve non-homogenous generalization for anonymizing on relational attributes while preserving the utility to the most degree. And we adopt grouping based on association rules for anonymizing on transaction attributes. Meanwhile, we also generate a graph of association rules which is non-trivial for improving data utility as a part of data publication. For the same privacy level, our approach preserves more useful information for satisfying different utility requirements via the combination of different anonymized techniques.

5 Experimental Study

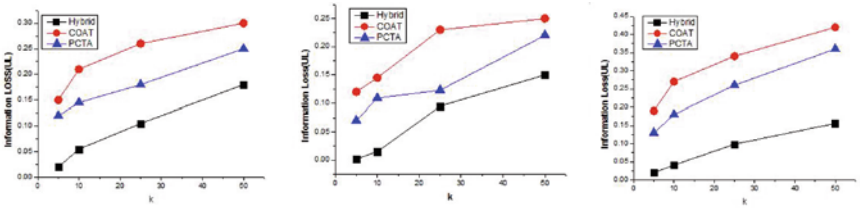
5.1 Experimental Framework

In this section, we present an extensive empirical evaluation of our privacy-preserving approach. We evaluate its utility, compare it to competing techniques. All experiments for this paper are implemented in C++ and SQL Server 2008. We use DBLP, BMS-WebView-2(BMS2) and BMS-POS (POS), three vastly different datasets that have been used in evaluating previous works [9,10,15,16,18] for ensuring the fairness of comparative experiments. These datasets are widely used as benchmark datasets in the knowledge discovery community. Their characteristics are listed in Table 1. In our experiment, L_v is synthetic data following the example of ADULT dataset since these datasets do not include relational attribute information.

We evaluate our approach as Hybrid against COAT [18], PCTA [16] and Safe (k,l)-grouping [15], in terms of data utility, under several different utility requirements. We compared the amount of data utility preserved by these methods by considering two utility measures: Utility Loss (*UL*) [16,18] and Expected Error (*ExpErr*) [15]. *UL* captures the utility loss by non-homogeneous generalization on relational attributes. And *ExpErr* captures the accuracy of query answering on anonymized data.

Table 1. Characteristics of the three datasets

Dataset	#Trans	#Distinct items	# Max.trans.size	# Avg.trans.size
DBLP	216753	170371	71	5.4
BMS-WebView-2	77370	3336	161	5.0
BMS-POS	306983	1177	5	2.65

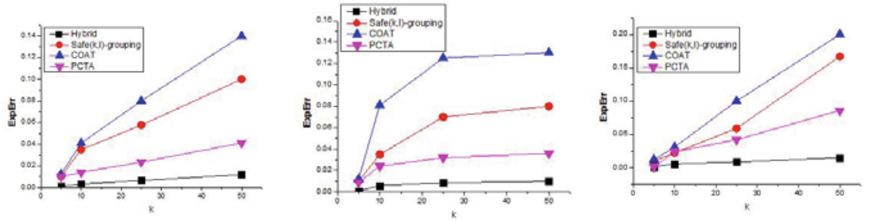


(a) *UL* vs. *k* on BMS2 (b) *UL* vs. *k* on POS (c) *UL* vs. *k* on DBLP

Fig. 4. Information loss on the three datasets

5.2 Experimental Results

Figure 4 plots the information loss under above approaches. Since Safe (k,l)-grouping is measured by aggregate query, we only compares the information loss of Hybrid to COAT and PCTA. As expected, increasing k induced more information loss due to the utility/privacy tradeoff. Our hybrid approach outperform than other approaches in all the three datasets. This is because, as k increases, the model of generalization in [16, 18] forces an increasingly large number of items to be generalized together, while our hybrid approach adopt grouping to perturb items instead of generalization. The impact of this generalization strategy on data utility was even more evident in the case of the DBLP dataset because of its big itemset, as shown in Fig. 4(c). This is expected, because generalization in [16, 18] will cause over-generalize when the size of itemset is "big", in which case substantial generalization is necessary. And our hybrid approach is stable than others.



(a) *ExpErr* vs. *k* on BMS2 (b) *ExpErr* vs. *k* on POS (c) *ExpErr* vs. *k* on DBLP

Fig. 5. The expected error on the three datasets

Figure 5 reports how well anonymized data supports query answering using *ExpErr*. As expected, increasing k induced more information loss because accurately answering queries involving many individuals is more difficult due to generalization. Observe that here, Hybrid are not raised obviously as k increases. This is because non-homogeneous generalization does not incur too much information distortion or information loss with the increases of k and the graph of association rules preserve the associations among items. So our approach can reduce information loss significantly for aggregate analysis. The quality of queries results shows that Hybrid proposed in this paper also consistently outperforms other approaches, using the expected error metric. Combined with the experimental results of Figs. 4 and 5, we can conclude that our hybrid approach can provide more accurate information for different applications than previous works employing the single anonymized technique.

6 Conclusion

In this paper, we introduced a hybrid optimization approach for anonymizing transactional data with relational and transaction attributes. To improve data utility, we anonymized relational attributes by adopting non-homogeneous generalization and anonymize transaction attributes via grouping based on the graph form. To satisfy different utility requirements, we constructed a graph of association rules as compensation to preserve the associations among items. Our experiment results demonstrated that our approach preserves data utility much better than previous works. The scalability of the proposed approach should be improved in our ongoing work.

Acknowledgment. The research is supported by the National Key Basic Research Program of China (973 Program, No. 2012CB326403), National Science Foundation of China (No. 61272535), Guangxi Bagui Scholar Teams for Innovation and Research Project, Guangxi Collaborative Innovation Center of Multi-source Information Integration and Intelligent Processing, Guangxi Natural Science Foundation (Nos. 2015GXNS FBA139246, 2013GXNSFBA019263, 2014GXNSF BA118288), Science and Technology Research Projects of Guangxi Higher Education (Nos. 2013YB029, 2015YB032), the Guangxi Science Research and Technology Development Project (No. 14124004-4-11) and Youth Scientific Research Foundation of Guangxi Normal University.

References

1. Chang, C.C., Thompson, B., Wang, H.W., Yao, D.: Towards publishing recommendation data with predictive anonymization. In: Proceedings of the 5th ACM Symposium on Information, Computer and Communications Security, pp. 24–35. ACM (2010)
2. Zheng, Z., Kohavi, R., Mason, L.: Real world performance of association rule algorithms. In: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 401–406. ACM (2001)

3. Xu, Y., Wang, K., Fu, A.W.C., Yu, P.S.: Anonymizing transaction databases for publication. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 767–775. ACM (2008)
4. Terrovitis, M., Mamoulis, N., Kalnis, P.: Privacy-preserving anonymization of set-valued data. Proc. VLDB Endowment **1**(1), 115–125 (2008)
5. Terrovitis, M., Mamoulis, N., Kalnis, P.: Local and global recoding methods for anonymizing set-valued data. VLDB J. Int. J. Very Large Data Bases **20**(1), 83–106 (2011)
6. He, Y., Naughton, J.F.: Anonymization of set-valued data via top-down, local generalization. Proc. VLDB Endowment **2**(1), 934–945 (2009)
7. Liu, J., Wang, K.: Anonymizing transaction data by integrating suppression and generalization. In: Zaki, M.J., Yu, J.X., Ravindran, B., Pudi, V. (eds.) PAKDD 2010, Part I. LNCS, vol. 6118, pp. 171–180. Springer, Heidelberg (2010)
8. Wang, L.E., Li, X.: A clustering-based bipartite graph privacy-preserving approach for sharing high-dimensional data. Int. J. Softw. Eng. Knowl. Eng. **24**(07), 1091–1111 (2014)
9. Ghinita, G., Tao, Y., Kalnis, P.: On the anonymization of sparse high-dimensional data. In: 2008 IEEE 24th International Conference on Data Engineering. ICDE 2008, pp. 715–724. IEEE (2008)
10. Ghinita, G., Kalnis, P., Tao, Y.: Anonymous publication of sensitive transactional data. IEEE Trans. Knowl. Data Eng. **23**(2), 161–174 (2011)
11. Wang, L., Li, X.: Personalized privacy protection for transactional data. In: Luo, X., Yu, J.X., Li, Z. (eds.) ADMA 2014. LNCS, vol. 8933, pp. 253–266. Springer, Heidelberg (2014)
12. Fung, B., Wang, K., Chen, R., Yu, P.S.: Privacy-preserving data publishing: a survey of recent developments. ACM Comput. Surv. (CSUR) **42**(4), 14 (2010)
13. Poulis, G., Loukides, G., Gkoulalas-Divanis, A., Skiadopoulos, S.: Anonymizing data with relational and transaction attributes. In: Blockeel, H., Kersting, K., Nijssen, S., Železný, F. (eds.) ECML PKDD 2013, Part III. LNCS, vol. 8190, pp. 353–369. Springer, Heidelberg (2013)
14. Takahashi, T., Sobataka, K., Takenouchi, T., Toyoda, Y., Mori, T., Kohro, T.: Top-down itemset recoding for releasing private complex data. In: 2013 Eleventh Annual International Conference on Privacy, Security and Trust (PST), pp. 373–376. IEEE (2013)
15. Cormode, G., Srivastava, D., Yu, T., Zhang, Q.: Anonymizing bipartite graph data using safe groupings. Proc. VLDB Endowment **1**(1), 833–844 (2008)
16. Gkoulalas-Divanis, A., Loukides, G.: Utility-guided clustering-based transaction data anonymization. Trans. Data Priv. **5**(1), 223–251 (2012)
17. Wong, W.K., Mamoulis, N., Cheung, D.W.L.: Non-homogeneous generalization in privacy preserving data publishing. In: Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, pp. 747–758. ACM (2010)
18. Loukides, G., Gkoulalas-Divanis, A., Malin, B.: Coat: constraint-based anonymization of transactions. Knowl. Inf. Syst. **28**(2), 251–282 (2011)
19. Gionis, A., Mazza, A., Tassa, T.: k-anonymization revisited. In: 2008 IEEE 24th International Conference on Data Engineering. ICDE 2008, pp. 744–753. IEEE (2008)
20. Karr, A.F., Kohnen, C.N., Oganian, A., Reiter, J.P., Sanil, A.P.: A framework for evaluating the utility of data altered to protect confidentiality. Am. Stat. **60**(3), 224–232 (2006)