Clustering-Based k-Anonymity

Xianmang He[1], HuaHui Chen[1], Yefang Chen[1],
Yihong Dong[1], Peng Wang[2], and Zhenhua Huang[3]

[1] School of Information Science and Technology, NingBo University
No.818, Fenghua Road, Ning Bo, 315122, P.R. China
{hexianmang,chenhuahui,chenyefang,dongyihong}@nbu.edu.cn
[2] School of Computer Science and Technology, Fudan University
No.220, Handan Road, Shanghai, 200433, P.R. China
pengwang5@fudan.edu.cn
[3] School of Electronic and Information Engineering, TongJi University
NO.1239. Siping Road, Shanghai, 200433, P.R. China
Huangzhenhua@tongji.edu.cn

Abstract. Privacy is one of major concerns when data containing sensitive information needs to be released for ad hoc analysis, which has attracted wide research interest on privacy-preserving data publishing in the past few years. One approach of strategy to anonymize data is generalization. In a typical generalization approach, tuples in a table was first divided into many QI (quasi-identifier)-groups such that the size of each QI-group is no less than k. Clustering is to partition the tuples into many clusters such that the points within a cluster are more similar to each other than points in different clusters. The two methods share a common feature: distribute the tuples into many small groups. Motivated by this observation, we propose a clustering-based k-anonymity algorithm, which achieves k-anonymity through clustering. Extensive experiments on real data sets are also conducted, showing that the utility has been improved by our approach.

Keywords: privacy preservation, algorithm, proximity privacy.

1 Introduction

Privacy leakage is one of major concerns when publishing data for statistical process or data analysis. In general, organizations need to release data that may contain sensitive information for the purposes of facilitating useful data analysis or research. For example, patients' medical records may be released by a hospital to aid the medical study. Records in Table 1 (called the microdata) is an example of patients' records published by hospitals. Note that attribute *Disease* contains sensitive information of patients. Hence, data publishers must ensure that no adversaries can accurately infer the disease of any patient. One straightforward approach to achieve this goal is excluding unique identifier attributes, such as *Name* from the table, which however is not sufficient for protecting privacy leakage under *linking-attack*[1, 2]. For example, the combination of *Age* and

P.-N. Tan et al. (Eds.): PAKDD 2012, Part I, LNAI 7301, pp. 405–417, 2012.

Table 1. Microdata T

	Age	Zip	Diesease
Andy	20	25	Flu
Bob	20	30	Bronchitis
Jane	30	25	Gastritis
Alex	40	30	Penumonia
Mary	50	10	Flu
Lily	60	5	Bronchitis
Lucy	60	10	Gastritis

Table 2. Generalization T^*

G-ID	Age	Zip	Diesease
1	[20-20]	[25-30]	Flu
1	[20-20]	[25-30]	Bronchitis
2	[30-40]	[25-30]	Gastritis
2	[30-40]	[25-30]	Penumonia
3	[50-60]	[5-10]	Flu
3	[50-60]	[5-10]	Bronchitis
3	[50-60]	[5-10]	Gastritis

Zipcode can be potentially used to identify an individual in Table 1, and has been called a quasi-identifier (QI for short)[1] in literatures. If an adversary has the background knowledge about Bob, that is: Age=20 and Zipcode=30, then by joining the background knowledge to Table 1, he can accurately infer Bob's disease, that is bronchitis.

To protect privacy against re-identifying individuals by joining multiple public data sources, k-anonymity ($k \geq 2$) was proposed, which requires that each record in a table is indistinguishable from at least $k-1$ other records with respect to certain quasi-identifiers. Generally, to achieve k-anonymity, generalization [1–3] is a popular methodology of privacy preservation for preventing linking attacks. Enough degree of generalization will hide a record in a crowd with at least k records with the same QI-values, thereby achieving k-anonymity. Table 2 demonstrates a generalized version of Table 1 (e.g., the Zip 30 of Bob, for instance, has been generalized to an interval [25, 30]). The generalization results in 3 equivalence classes, as indicated by their group-IDs. Each equivalence class is referred to as a QI-group. As a result, given Table 2, even if an adversary has the exact QI-values of Bob, s/he still can not exactly figure out the tuple of Bob from the first QI-group.

1.1 Motivation

Although generalization-based algorithms have successfully achieved the privacy protection objective, as another key issue in data anonymization *utility* still needs to be carefully addressed. Great efforts have been dedicated to developing algorithms that improve utility of anonymized data while ensuring enough privacy-preservation. One of the direct measures of the utility of the generalized data is information loss. In order to make the anonymized data as useful as possible for certain applications, it is required to reduce the information loss as much as possible. In general, *the less total information loss leads to better utility, which reflects its usefulness as one of the steps in exploratory data analysis.*

Clustering [4] is a method commonly used to automatically partition a data set into many groups. As an example of clustering is depicted in Figure 1-3. The input points are shown in Figure 1, and the steps to the desired clusters are shown in Figure 2 and Figure 3. Here, points belonging to the same cluster are given the same color.

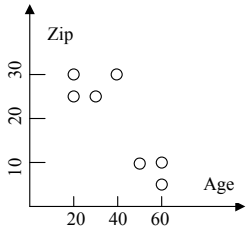


Fig. 1. The points inTable 1

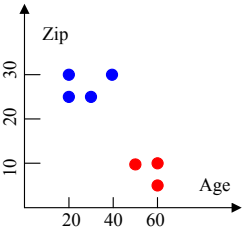


Fig. 2. Data Clustering (Step 1)

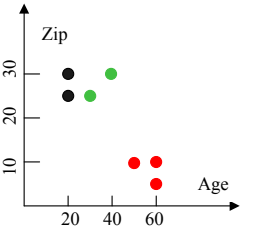


Fig. 3. Data Clustering (Step 2)

Then, we may wonder: *Can we significantly improve the utility while preserving k-anonymity by clustering-based approaches ?* The answer depends on whether it is possible to partition microdata into clusters with less information loss while still ensuring k-anonymity. Intuitively, data points within a cluster are more similar to each other than they are to a point belonging to a different cluster.

The above observation motivates us to devise a new solution to improve the data utility of clustering-based solutions. As an example, we illustrate the details to generalize Table 1 by our approach. Let *gen* be a generalization function that takes as input a set of tuples and returns a generalized domain. Firstly, Table 1 is divided into 2 clusters, denoted by red and blue in Figure 2, respectively. Then, the cluster denoted by blue is further divided into 2 cluster, denoted by black and green color in Figure 3. Finally, tuples with same color are generalized as a QI-group, that is, tuple Andy and Bob consists of the first QI-group, and assign gen({Andy, Bob})=$\langle[20-20],[25-30]\rangle$ to the first QI-group. Similarly, {Jane,Alex}, {Mary, Lily, Lucy} make the second and third QI-group. Eventually, table 2 is the final result by our approach.

In this paper, we mainly focused on the basic k-anonymity model due to the following reasons: (i) k-anonymity is a fundamental model for privacy protection, which has received wide attention in the literatures; (ii) k-anonymity has been employed in many real applications such as location-based services [5, 6], where there are no additional (sensitive) attributes; (iii) There is no algorithm that is suitable for so many privacy metrics such as l-diversity[7], t-Closeness [8], but algorithms for k-anonymity are simple yet effective, and can be further adopted for other privacy metrics. Apart from the k-anonymity model, we also consider the scenarios with stronger adversaries, extending our approach to l-diversity(Section 4)

The rest of the paper is organized as follows. In Section 2, we give the definitions of basic concept and the problem will be addressed in this paper. In Section 3, we present the details of our generalization algorithm. Section 4 discusses the extension of our methodology for l-diversity. We review the previously related research in Section 5. In Section 6, we experimentally evaluate the efficiency and effectiveness of our techniques. Finally, the paper is concluded in Section 7.

2 Fundamental Definitions

Let T be a microdata table that contains the private information of a set of individuals and has d QI-attributes $A_1, ..., A_d$, and a sensitive attribute A_s. We consider that A_s is numerical, and every QI-attribute $A_i (1 \leq i \leq d)$ can be either numerical or categorical. All attributes have finite and positive domains. For each tuple $t \in T, t.A_i (1 \leq i \leq d)$ denotes its value on A_i, and $t.A_s$ represents its SA value.

2.1 Basic Concept

A *quasi-identifier* $QI = \{A_1, A_2, \cdots, A_d\} \subseteq \{A_1, A_2, \cdots, A_n\}$ is a minimal set of attributes, which can be joined with external information in order to reveal the personal identity of individual records.

A *partition* P consists of several subsets $G_i (1 \leq i \leq m)$ of T, such that each tuple in T belongs to exactly one subset and $T = \bigcup_i^m G_i$. We refer to each subset G_i as a QI-group.

2.2 *K*-means Clustering

K-means clustering [4] is a method commonly used to automatically partition a data set into K groups. It proceeds by selecting K initial cluster centers and then iteratively refining them as follows:

Step 1. Each tuple t_i is assigned to its closest cluster center.

Step 2. Each cluster center C_j is updated to be the mean of its constituent instances.

The algorithm converges when there is no further change in assignment of tuples to clusters. In this work, we initialize the clusters using instances picked at random from the data set. The data sets we used are composed solely of either numeric features or categorical features. For both numeric and categorical features, we adopt the normalized certainty penalty(see the definition 1) to measure the distance.

The final issue is how to choose K. To keep the algorithm simple in this paper, we consider binary partitioning, that is, K is fixed as 2.

2.3 Problem Definition

Some methods have been developed to measure the information loss in anonymization. In this paper, we adopt the normalized certainty penalty to measure the information loss.

Definition 1 (Normalized Certainty Penalty [9]). *Suppose a table T is anonymized to T^*. In the domain of each attribute in T, suppose there exists a*

global order on all possible values in the domain. If a tuple t in T^* has range $[x_i, y_i]$ on attribute $A_i (1 \leq i \leq d)$, then the normalized certainty penalty in t on A_i is $NCP_{A_i}(t) = \frac{|y_i - x_i|}{|A_i|}$, where $|A_i|$ is the domain of the attribute A_i. For tuple t, the normalized certainty penalty in t is $NCP(t) = \sum_i^d w_i \cdot NCP_{A_i}(t)$, where w_i is the weight of attribute A_i. The normalized certainty penalty in T is $\sum_{t \in T^*} NCP(t)$.

Now, we are ready to give the formal definition about the problem that will be addressed in this paper. Information loss is an unfortunate consequence of data anonymization. We aim to generate a utility-friendly version anonymizaiton for a microdata such that the privacy can be guaranteed by k-anonymity and the information loss quantified by NCP is minimized. Now, we are ready to give the formal definition about the problem that will be addressed in this paper. (Limited by space, all proofs are omitted.)

Definition 2 (Problem Definition). *Given a table T and an integer k, anonymize it by clustering to be T^* such that T^* is k-anonymity and the total information loss is minimized measured by NCP.*

Theorem 1. *(Complexity) The problem of optimal clustering-based anonymization is NP-hard under the metric NCP.*

3 Clustering-Based Generalization Algorithm

In this section, we will present the details of our clustering-based anonymization approach. The key of our algorithm is to divide all tuples into more compact clusters efficiently and correctly. We now proceed to a discussion of our modifications to the K-means algorithm.

To keep the algorithm simple, we consider binary clustering. That is, in each round, we partition a set of tuples into two subsets by clustering. In order to reduce the total information loss, we will cluster the microdata following the idea: *distribute tuples sharing the same or quite similar QI-attributes into the same cluster*. We adopt the NCP to measure the distance. The detailed partitioning procedure is presented in Figure 4. Initially, S contains T itself (line 1); then, each $G \in S$ is divided into two generalizable subsets G_1 and G_2 such that $G_1 \cup G_2 = G$, $G_1 \cap G_2 = \emptyset$ (line 5-7).

The size of the two subsets should $\geq k$, otherwise adjustment is needed (line 8). Without loss of generality, assume that $G_1 < k$, we need to borrow $k - |G_1|$ tuples from G_2 to make sure that G_1 has a cardinality $\geq k$.

The tries to converges will cost unacceptable time, to accelerate the partitioning, the attempts to cluster G are tried r times and tuples of G are randomly shuffled for each time (line 4). Our experimental results show that most of G can be partitioned into two sub-tables by up to $k = 15$ tries. The algorithm stops when no sub-tables in S can be further partitioned.

By the Lemma in the paper [9, 10] that the optimal k-anonymity partitioning of microdata does not contain groups of more than $2k - 1$ records, we have that

the partitioning algorithm will terminate when the size of all groups is between k and $2k-1$. If at least one group contains a cardinality more than $2k-1$, the partitioning algorithm will continue.

In the above procedure, the way that we partition G into two subsets G_1 and G_2 is influential on the information loss of the resulting solution. In the first round, we randomly choose two tuples t_1, t_2 as the center points C_1, C_2 , and then insert them G_1 and G_2 separately. Then, we distribute each tuple $w \in G$: for each tuple w, we compute $\Delta_1 = NCP(C_1 \cup w)$ and $\Delta_2 = NCP(G_2 \cup w)$, and add tuple w to the group that leads to lower penalty (line 7). If $\Delta_1 = \Delta_2$, assign the tuple to the group who has lower cardinality. After successfully partitioning G, remove the tuples t_1 and t_2 from $G_1 - \{t_1\}$ and $G_2 - \{t_2\}$. At the later each round, the center points C_i are conducted as follows: $C_i = \frac{\sum_{t \in G_i} t}{|G_i|}, i = 1, 2.$ that is, for each attribute $A_j (1 \leq j \leq d)$, $C_i.A_j = \frac{\sum_{t \in G_i} t.A_j}{|G_i|}, i = 1, 2.$

After the each partition, if the current partition is better than previous tries, record the partition result G_1, G_2 and the total sum of $NCP(G_1)$ and $NCP(G_2)$. That is, we pick the one that that minimizes the sum of $NCP(G_1)$ and $NCP(G_2)$ as the final partition among the r partitions(line 9). Each round of G can be accomplished in $\mathrm{O}(r \cdot (|G| \cdot (6 + \lambda)))$ expected time, where λ is the cost of evaluating loss. The computation cost is theoretically bounded in Theorem 2.

Theorem 2. *For microdata T, the clustering-based algorithm can be accomplished in $O(r \cdot |T| \cdot log(|T|))$ average time, where r is the number of rounds , and $|T|$ is the cardinality of microdata T.*

Input: A microdata T, integers k and rounds r
Output: anonymized table T^*;
Method:
/* the parameter r is number of rounds to cluster G^*/
1. $S = \{T\}$;
2. While($\exists G \in S$ that $|G| \geq 2k$)
3. For $i = 1$ to r
4. Randomly shuffle the tuples of G;
5. Set Center $C_i = \frac{\sum_{t \in G_i} t}{|G_i|}$;
6. Set $G_1 = G_2 = \emptyset$;
7. Distribute each tuple w in G:
compute $\Delta_1 = NCP(w \cup C_1)$ and $\Delta_2 = NCP(w \cup C_2)$;
If($\Delta_1 < \Delta_2$) then add w to G_1, else add w to G_2;
8. Adjust G_1, G_2 that each group has at least k tuples;
9. If the current partition is better than previous tries, record G_1 and G_2;
10. Remove G from S, and add G_1, G_2 to S;
11. Return S;

Fig. 4. The partitioning algorithm

4 Extension to l-Diversity

In this section, we discuss how we can apply clustering-based anonymization for other privacy principles. In particular, we focus on l-diversity, described in Definition 3.

Definition 3 (l-diversity[7]). *A generalized table T^* is l-diversity if each QI-group $QI_i \in T^*$ satisfies the following condition: let v be the most frequent A_s value in QI_i, and $c_i(v)$ be the number of tuples $t \in QI_i$, then $\frac{c_i(v)}{|QI_i|} \leq \frac{1}{l}$.*

To generalize a table through clustering-based anonymization, we partition a table into sub-tables T_i which satisfy l-diversity: after each round of the above partitioning, if both (G_1 and G_2)) satisfy l-diversity, we remove G from S, and add G_1, G_2 to S; otherwise G is retained in S. Then for each subset $T_i \in S$, we conduct the splitting algorithm (see Figure 5) to produce the final l-diverse partitions.

The principle l-diversity demands that: the number of the most frequent A_s value in each QI-group QI_i can't exceed $\frac{|QI_i|}{l}$. Motivated by this, we arrange the tuples to a list ordered by its A_s values, then distribute the tuples in L into $QI_i(1 \leq i \leq g)$ a round-robin fashion. The resulting splitting is guaranteed to be l-diversity, which is stated in Theorem 3. (If table T with sensitive attribute A_s satisfies $\max\{c(v) : v \in T.A_s\} > \frac{|T|}{l}$, then there exists no partition that is l-diversity.)

Input: table T, parameter l
Output: QI-groups QI_j that satisfy l-diversity;
Method:
1. If $\max\{c(v) : v \in T.A_s\} \geq \frac{|T|}{l}$, Return;
2. Hash the tuples in T into groups $Q_1, Q_2, \cdots, Q_\lambda$ by their A_s values;
3. Insert these groups $Q_1, Q_2, \cdots, Q_\lambda$ into a list L in order;
4. Let $g = \frac{|T|}{l}$, set QI-groups $QI_1 = QI_2 = \cdots = QI_g = \emptyset$;
5. Assign tuple $t_i \in L$ $(1 \leq i \leq |L|)$ to QI_j, where $j = (i \bmod g) + 1$

Fig. 5. The splitting algorithm

Theorem 3. *If table T with sensitive attribute A_s satisfies* $\max\{c(v) : v \in T.A_s\} \leq \frac{|T|}{l}$ *(where $c(v)$ is the number of tuples in T with sensitive value v), the partition produced by our splitting algorithm fulfills l-diversity.*

5 Related Work

In this section, previous related work will be surveyed. Existing generalization algorithms can be further divided into heuristic-based and theoretical-based approaches. Generally, appropriate heuristics are general so that they can be used

Table 3. Summary of attributes

Attribute	Number of distinct values	Types
Age	78	Numerical
Gender	2	Categorical
Education	17	Numerical
Marital	6	Categorical
Race	9	Numerical
Work-class	10	Categorical
Country	83	Numerical
Occupation	50	Sensitive

(a) SAL

Attribute	Number of distinct values	Types
Age	78	Numerical
Occupation	711	Numerical
Birthplace	983	Numerical
Gender	2	Categorical
Education	17	Categorical
Race	9	Categorical
Work-class	9	Categorical
Marital	6	Categorical
Income	[1k,10k]	Sensitive

(b) INCOME

Table 4. Parameters and tested values

Parameter Values	
k	250,200,150,**100**,50
cardinality n	**100k**,200k,300k,400k,500k
number of QI-attributes d	3,4,5,**6**

in many anonymization models. To reduce information loss, efficient greedy solutions following certain heuristics have been proposed [9–13] to obtain a near optimal solution. Generally, these heuristics are general enough to be used in many anonymization models. Incognito [14] provides a practical framework for implementing full-domain generalization, borrowing ideas from frequent item set mining, while [10] presents a framework mapping the multi-dimensional quasi-identifiers to 1-Dimensional(1-D) space. For 1-D quasi-identifiers, an algorithm of $O(K \cdot N)$ time complexity for optimal solution is also developed. It is discovered that k-anonymizing a data set is strikingly similar to building a spatial index over the data set, so that classical spatial indexing techniques can be used for anonymization [15]. To achieve k-anonymity, Mondrian [16] takes a partitioning approach reminiscent of KD-trees.

The idea of non-homogeneous generalization was first introduced in [17], which studies techniques with a guarantee that an adversary cannot associate a generalized tuple to less than K individuals, but suffering additional types of attack. Authors of paper [13] proposed a randomization method that prevents such type of attack and showed that k-anonymity is not compromised by it, but its partitioning algorithm is only a special of the top-down algorithm presented in [9]. The model of the paper [13, 17], the size of QI-groups is fixed as 1.

The algorithms mentioned above work well on practical data sets, but do not have attractive asymptotical performance in the worst case. This motivates studies on the theoretical aspects of k-anonymity [16, 18]. Most of these works show that the problem of optimal k-anonymity is NP-hard even a simple quality metric is employed.

6 Empirical Evaluation

In this section, we will experimentally evaluate the effectiveness and efficiency of the proposed techniques. Specifically, we will show that by our technique (presented in Section 3) have significantly improved the utility of the anonymized data with quite small computation cost.

Towards this purpose, two widely-used real databases sets: SAL and INCOME(downloadable from http://ipums.org) with 500k and 600k tuples, respectively, will be used in following experiments. Each tuple describes the personal information of an American. The two data sets are summarized in Table 3.

In the following experiments, we compare our cluster-based anonymity algorithm (denoted by CB) with the existing state-of-the-art technique: the non-homogeneous generalization [13](NH for short). (The fast algorithm [10] was cited and compared with NH in the paper [13], therefore, we omit the details of the fast algorithm.)

In order to explore the influence of dimensionality, we create two sets of micro-data tables from SAL and INCOME. The first set has 4 tables, denoted as SAL-3, $\cdots$, SAL-6, respectively. Each SAL-d ($3 \leq d \leq 6$) has the first d attributes in Table 3 as its QI-attributes and Occupation as its sensitive attribute(SA). For example, SAL-4 is 5-Dimensional, and contains QI-attributes: Age, Gender, and

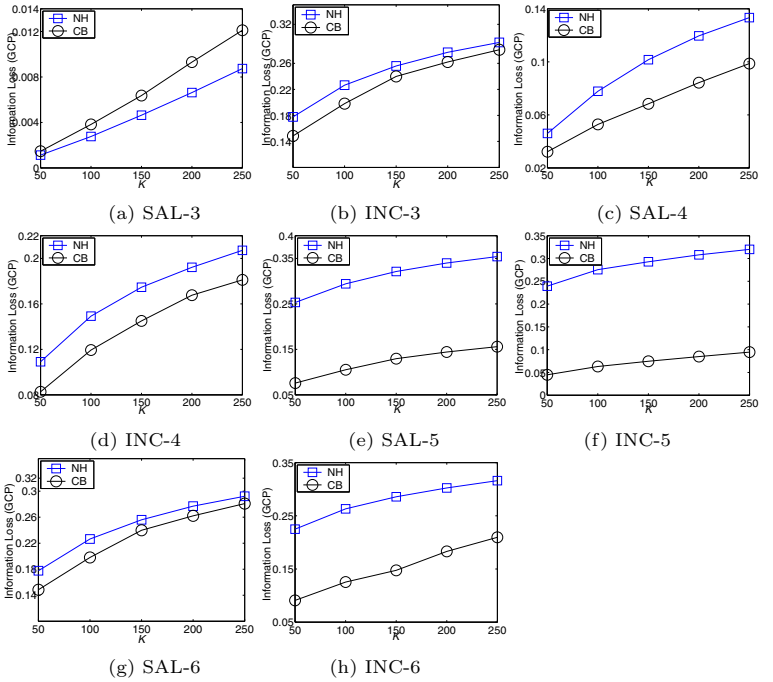


Fig. 6. The Global Certainty Penalty vs. parameters k and d

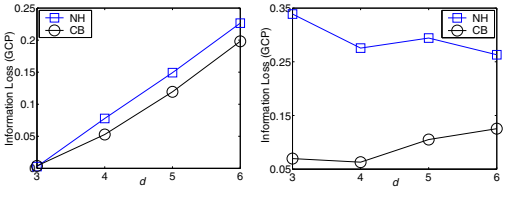
Education, Marital. The second set also has 4 tables INC-3, $\cdots$, INC-6, where each INC-d ($3 \leq d \leq 6$) has the first d attributes as QI-attributes and income as the SA.

In the experiments, we investigate the influence of the following parameters on information loss of our approach: (i) value of k in k-anonymity; (ii)number of attributes d in the QI-attributes; (iii)number of tuples n. Table 4 summarizes the parameters of our experiments, as well as their values examined. **Default values are in bold font**. Data sets with different cardinalities n are also generated by randomly sampling n tuples from the *full* SAL-d or INC-d ($3 \leq d \leq 6$). All experiments are conducted on a PC with 1.9 GHz AMD Dual Core CPU and 1 gigabytes memory. All the algorithms are implemented with VC++ 2008.

We measure the information loss of the generalized tables using GCP, which is first used in [10]. Note that GCP essentially is equivalent to NCP with only a difference of constant number $d \times N$. Specifically, under the same partition P of table T, $GCP(T) = \frac{NCP(T)}{d \times N}$(d is the size of QI-attributes), when all the weights are set to 1.0.

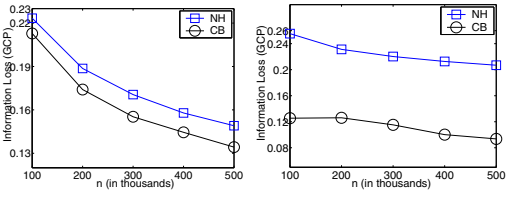
6.1 Privacy Level *K*

In order to study the influence of k on data utility, we observe the evolution of GCP that has been widely used to measure the information loss of the generalized tables by varying k from 50 to 250 with the increment of 50. In all following experiments, without explicit statements, default values in Table 4 will be used for all other parameters. The results on SAL-d and INC-d ($3 \leq d \leq 6$) data are



(a) SAL-d (b) INC-d

Fig. 7. The Global Certainty Penalty vs. QI-Dimensionality d



(a) SAL-7 (b) INC-7

Fig. 8. The Global Certainty Penalty vs. Cardinality n

shown in Figure 6 (a)-6(h). From the results, we can clearly see that information loss of CB sustains a big improvement over NH, for the tested data except the on SAL-3. Another advantage of our model over NH is that the utility achieved by our model is less sensitive to domain size than NH. From the figures, we can see that data sets generated by NH has a lower GCP on SAL-d than that on INC-d ($4 \leq d \leq 7$) due to the fact that domain size of SAL is smaller than that of INC. Such a fact implies that the information loss of NH is positively correlated to the domain size. However, in our model, domain size of different data set has less influence on the information loss of the anonymized data.

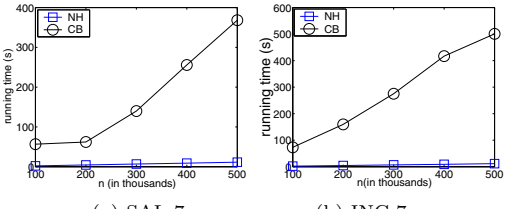
Results of this experiment also suggest that for almost all tested data sets the GCP of these algorithms grows linearly with k. This can be reasonably explained since larger k will lead to more generalized QI-groups, which inevitably will sacrifice data utility. NH performs well when the dimensionality of QI-Attributes is low and the domain size is small, see the experiment results in the paper[13].

6.2 QI-Attributes Dimensionality d

Experiments of this subsection is designed to show the relation between the information loss of these algorithms and data dimensions d. In general, the information loss will increase with d, since data sparsity or more specifically the data space characterized by a set of attributes exponentially increases with the number of attributes in the set, i,e, dimensions of the table. Figure 7(a) and 7(b) compare the information loss of the anonymization generated by the these four methods with respect to different values of d on SAL-d and INC-d, respectively. It is clear that the anonymization generated by the cluster-based method has a lower global certainty penalty compared to that of NH. The advantage of CB is obvious, and such an advantage of CB can be consistently achieved when d lies between 4 to 6.

6.3 Cardinality of Data Set n

In this subsection, we investigate the influence of the the table size n on information loss. The results of experiments on two data sets SAL-7 and INC-7 are shown in Figure 8(a) and 8(b), respectively. We can see that the information



(a) SAL-7 (b) INC-7

Fig. 9. Running time vs. Cardinality n

loss of these methods on both two data sets decreases with the growth of n. This observation can be attributed to the fact that when the table size increases more tuples will share the same or quite similar QI-attributes. As a result, it is easier for the partitioning strategies to find very similar tuples to generalize. Similar to previously experimental results, our method is the clear winner since information loss of CB is significantly small than that of NH, which is consistently observed for various database size.

6.4 Efficiency

Finally, we evaluate the overhead of performing anonymization. Figure 9(a) and 9(b) show the computation cost of the these anonymization methods on two data sets, respectively. We compare CB with NH when evaluating computational cost. The running time of tow algorithms increases linearly when n grows from 100k to 500k, which is expected since more tuples that need to be anonymized will cost longer time to finish the anonymization procedure. The NH method is more efficient. Comparison results show that the advantages of our method in anonymization quality do not come for free. However, in the worst case, our algorithm can be finished in 500 seconds, which is acceptable. In most real applications quality is more important than running time, which justifies the strategy to sacrifice certain degree of time performance to achieve higher data utility.

Summary. Above results clearly show that clustering-based anonymization achieves less information loss than the non-homogeneous anonymization (NH) in cases where the dimensionality of QI-attribute $d > 3$. NH has a good performance when the domain size is small, and the dimensionality of QI-Attributes is low. This is due to its greedy partitioning algorithm.

7 Conclusion

As privacy becomes a more and more serious concern in applications involving microdata, good anonymization is of significance. In this paper, we propose an algorithm which is based on clustering to produce a utility-friendly anonymized version of microdata. Our extensive performance study shows that our methods outperform the non-homogeneous technique where the size of QI-attribute is larger than 3.

Acknowledgement. This work was supported in part by the National Natural Science Foundation of China (NO.6090303, NO.60902097, NO.60973047), the Natural Science Foundation of Zhejiang Province of China under Grant No. Y1091189 and No.Y1090571, the Research Fund for the Doctoral Program of Higher Education(No. 20090072120056), the Special Fund for Information Development of Shanghai (No. 200901015), the National High-Tech Research and Development Plan of China (No. 2008AA04Z106), the Project of Science and Technology Commission of Shanghai Municipality (NO.08DZ1122300), and the Major Scientific & Technology Specific Programs of Zhejiang Province for Key Industrial Project(No.2011C11042).

References

1. Sweeney, L.: k-anonymity: a model for protecting privacy. Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 10(5), 557–570 (2002)
2. Samarati, P.: Protecting respondents' identities in microdata release. TKDE 13(6), 1010–1027 (2001)
3. Samarati, P., Sweeney, L.: Generalizing data to provide anonymity when disclosing information. In: PODS 1998, p. 188. ACM, New York (1998)
4. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations, Berkeley, pp. 281–297 (1967)
5. Kalnis, P., Ghinita, G., Mouratidis, K., Papadias, D.: Preventing location-based identity inference in anonymous spatial queries. TKDE 19(12), 1719–1733 (2007)
6. Mokbel, M.F., Chow, C.-Y., Aref, W.G.: The new casper: query processing for location services without compromising privacy. In: VLDB 2006, pp. 763–774 (2006)
7. Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramaniam, M.: l-diversity: Privacy beyond k-anonymity. In: ICDE 2006, p. 24 (2006)
8. Li, N., Li, T.: t-closeness: Privacy beyond k-anonymity and l-diversity. In: KDD 2007, pp. 106–115 (2007)
9. Xu, J., Wang, W., Pei, J., Wang, X., Shi, B., Fu, A.W.-C.: Utility-based anonymization using local recoding. In: KDD 2006, pp. 785–790. ACM (2006)
10. Ghinita, G., Karras, P., Kalnis, P., Mamoulis, N.: Fast data anonymization with low information loss. In: VLDB 2007, pp. 758–769. VLDB Endowment (2007)
11. Fung, B.C.M., Wang, K., Yu, P.S.: Top-down specialization for information and privacy preservation, pp. 205–216 (2005)
12. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Workload-aware anonymization. In: KDD 2006, pp. 277–286. ACM, New York (2006)
13. Wong, W.K., Mamoulis, N., Cheung, D.W.L.: Non-homogeneous generalization in privacy preserving data publishing. In: SIGMOD 2010, pp. 747–758. ACM, New York (2010)
14. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Incognito: efficient full-domain k-anonymity. In: SIGMOD 2005, pp. 49–60. ACM, New York (2005)
15. Iwuchukwu, T., Naughton, J.F.: K-anonymization as spatial indexing: toward scalable and incremental anonymization. In: VLDB 2007, pp. 746–757 (2007)
16. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Mondrian multidimensional k-anonymity. In: ICDE 2006, Washington, DC, USA, p. 25 (2006)
17. Gionis, A., Mazza, A., Tassa, T.: k-anonymization revisited. In: ICDE 2008, pp. 744–753. IEEE Computer Society, Washington, DC (2008)
18. Bayardo, R.J., Agrawal, R.: Data privacy through optimal k-anonymization. In: ICDE 2005, pp. 217–228. IEEE Computer Society, Washington, DC (2005)