Priority Driven K-Anonymisation for Privacy Protection

Xiaoxun Sun[1] **Hua Wang**[1] **Jiuyong Li**[2]

[1] Department of Mathematics & Computing
University of Southern Queensland
Toowoomba, Queensland 4350, Australia
Email: {sunx, wang}@usq.edu.au

[2] School of Computer and Information Science
University of South Australia, Adelaide, Australia
Email: jiuyong.li@unisa.edu.au

Abstract

Given the threat of re-identification in our growing digital society, guaranteeing privacy while providing worthwhile data for knowledge discovery has become a difficult problem. k-anonymity is a major technique used to ensure privacy by generalizing and suppressing attributes and has been the focus of intense research in the last few years. However, data modification techniques like generalization may produce anonymous data unusable for medical studies because some attributes become too coarse-grained. In this paper, we propose a priority driven k-anonymisation that allows to specify the degree of acceptable distortion for each attribute separately. We also define some appropriate metrics to measure the distance and information loss, which are suitable for both numerical and categorical attributes. Further, we formulate the priority driven k-anonymisation as the k-nearest neighbor (KNN) clustering problem by adding a constraint that each cluster contains at least k tuples. We develop an efficient algorithm for priority driven k-anonymisation. Experimental results show that the proposed technique causes significantly less distortions.

Keywords: K-Anonymity; Privacy Protection;

1 Introduction

Agencies and other organizations often need to publish microdata, e.g. medical data or census data, for research and other purpose. However, if individuals can be uniquely identified in the microdata then their private information (such as their medical condition) would be disclosed, and this is unacceptable. To avoid the identification of records in microdata, uniquely identifying information like names and social security numbers are removed from the table. Unfortunately, simply removing unique identifiers (e.g., names or phone numbers) from data is not enough, as individuals can still be identified when external data is linked to de-identified data, by using a combination of non-unique attributes such as age and postcode. Such non-unique attributes are often called quasi-identifiers (QIDs).

A recent study estimated that 87% of the population of the United States can be uniquely identified "linking attack" using the seemingly innocuous attributes gender, date of birth, and 5-digit zip code (Sweeney 2000). To avoid linking attacks, Samarati and Sweeney (Samarati 2001, Sweeney 2002a) proposed a privacy principle called k-anonymity. It works by replacing a QID value with a more general one, such that the generalized QID values of each tuple are made identical to at least $k-1$ other tuples in the anonymized table. This generalization process trades-off data utility for privacy protection. To illustrate this, consider Tables 1. Table 1(c) is a possible 2-anonymous view of Table 1(a). Here, queries such as "how many people live in an area with a postcode between 4350 and 4353 are male?" can no longer be answered accurately, and it is also more difficult to infer sensitive disease information about the individuals contained in the table.

Although the idea of k-anonymity is conceptually straightforward, the computational complexity of finding an optimal solution for the k-anonymity problem has been shown to be NP-hard, even when one considers only cell suppression (Aggarwal et al. 2005, Meyerson &Williams 2004, Sun et al. 2008b). The k-anonymity problem has recently drawn considerable interest from research community, and a number of algorithms have been proposed (Bayardo et al. 2005, Fung et al. 2005, Leferve et al. 2005, Sweeney 2002b, Sun et al. 2008a). Current solutions, however, suffer from high cost of information loss mainly due to relying on pre-defined generalization hierarchies (Fung et al. 2005, Leferve et al. 2005, Sweeney 2002b, Sun et al. 2008a) or total order imposed on each attribute domain (Bayardo et al. 2005). A more general view of k-anonymisation is clustering with a constraint of the minimum number of objects in every cluster (Aggarwal et al. 2006, Byun et al. 2006). A number of methods approach identity protection by clustering (Agrawal 2001, Aggarwal 2005). However, these methods are applicable to numerical attributes only. A recent work (Domingo-Ferrer et al. 2005) extends a clustering-based method (Domingo-Ferrer et al. 2002) to ordinal attributes, but it does not deal with attributes in hierarchical structures.

In this paper, we propose a priority driven k-anonymisation that allows to specify the degree of acceptable distortion for each attribute separately. We also define some appropriate metrics to measure the distance and information loss, which are suitable for both numerical and categorical attributes. Further, we formulate the priority driven k-anonymisation as the k-nearest neighbor (KNN) clustering problem by adding a constraint that each cluster contains at least k tuples. We develop an efficient algorithm for priority driven k-anonymisation. Experimental results show that the proposed technique causes significantly less distortions.

Copyright©2008, Australian Computer Society, Inc. This paper appeared at conference Seventh Australasian Data Mining Conference (AusDM 2008), Glenelg, Australia. Conferences in Research and Practice in Information Technology, Vol. 87. John F. Roddick, Jiuyong Li, Peter Christen and Paul Kennedy, Eds. Reproduction for academic, not-for profit purposes permitted provided this text is included.

Gender	Age	Pcode	Problem
male	middle	4350	stress
male	middle	4350	obesity
male	young	**4351**	stress
female	young	**4352**	obesity
female	old	4353	stress
female	old	4353	obesity

Gender	Age	Pcode	Problem
male	middle	4350	stress
male	middle	4350	obesity
*****	young	**435***	stress
*****	young	**435***	obesity
female	old	4353	stress
female	old	4353	obesity

Gender	Age	Pcode	Problem
*	middle	435*	stress
*	middle	435*	obesity
*	young	435*	stress
*	young	435*	obesity
*	old	435*	stress
*	old	435*	obesity

Table 1: (a) Left: a raw table. (b) Middle: a 2-anonymous table by local recoding. (c) Right: a 2-anonymous view by global recoding.

2 Preliminary Definitions

The objective of k-anonymisation is to make every tuple of privacy-related attributes in a published table identical to at least $k-1$ other tuples. As a result, no privacy-related information can be easily inferred. For example, young people with stress and obesity are potentially identifiable by their unique combinations of gender, age and postcode attributes in Table 1(a). To preserve their privacy, we may generalize their gender and postcode attribute values such that each tuple in attribute set {Gender, Age, Postcode} has two occurrences. The view after the generalization is listed in Table 1(b).

Definition 1 *A quasi-identifier(QID) attribute set is a set of attributes in a table that potentially reveal private information, possibly by joining with other tables. A QI-group of a table with respect to the QID attribute set is the set of all tuples in the table containing identical values for the QID attribute set.*

For example, the attribute set {Gender, Age, Postcode} in Table 1(a) is a QID and Tuples 1 and 2 in Table 1(b) form a QI-group with respect to this QID since their corresponding values are identical. Table 1(a) potentially reveals private information of patients (e.g. young patients with stress and obesity). If the table is joined with other tables, it may reveal more information of patients' disease history. Normally, the QID set is understood by domain experts.

Definition 2 (k-anonymity) *A table is called k-anonymous with respect to a QID if the size of every QI-group with respect to the QID set is at least k.*

k-anonymity requires that every occurrence within an attribute set has the frequency at least k. For example, Table 1(a) does not satisfy 2-anonymity property since tuples male, young, 4351 and female, young, 4352 occur only once. Table 1(b) is a 2-anonymous view of Table 1(a) since the size of all QI-group with respect to the QID is 2.

Another objective for k-anonymisation is to minimize distortions. A table may have more than one k-anonymous views, but some are better than others. For example, we may have another 2-anonymous view of Table 1(a) as in Table 1(c). Table 1(c) loses much more information than Table 1(b).

In the literature of k-anonymity problem, there are two main models. One model is global recoding (Fung et al. 2005, Leferve et al. 2005, Sweeney 2002a, Samarati 2001), while the other is local recoding (Aggarwal et al. 2005, Sweeney 2002b). Here, we assume that each attribute has a corresponding conceptual generalization hierarchy or taxonomy tree. A lower level domain in the hierarchy provides more details than a higher level domain. For example, Postcode 4350 is a lower level domain and Postcode 435* is a higher level domain. We assume such hierarchies for numerical attributes too. In particular, we have a hierarchical structure defined with {value, interval, *}, where value is the raw numerical data, interval is the range of the raw data and * is a symbol representing any values. Generalization replaces lower level domain values with higher level domain values. For example, Age 27, 28 in the lower level can be replaced by the interval (27-28) in the higher level. Examples of global and local recoding are shown in Table 1(b) and Table 1(c).

Definition 3 ((Li et al. 2006)) *Let h be the height of a domain hierarchy, and let levels 1, 2, $\cdots$, $h-1$, h be the domain levels from the most general to most specific, respectively. Let the weight between domain level $j-1$ and j be predefined, denoted by $w_{j,j-1}$, where $2 \leq j \leq h$. When a cell is generalized from level p to level q, where $p > q$. The weighted hierarchical distance of this generalization is defined as:*

$$WHD(p,q) = \frac{\sum_{j=q+1}^{p} w_{j,j-1}}{\sum_{j=2}^{h} w_{j,j-1}}$$

In the following, we discuss two simple but typical schemes to define $w_{j,j-1}$.

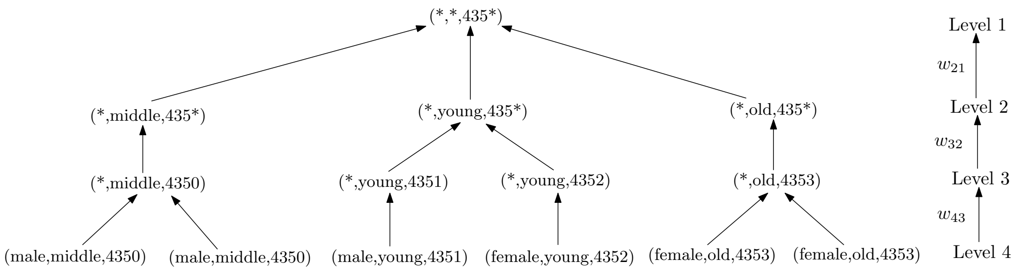
(1). Uniform Weight: $w_{j,j-1} = 1$ $(2 \leq j \leq h)$

This is the simplest scheme where all weights are equal to 1. In this scheme, WHD is the number of steps a cell being generalized over all possible generalization steps. For example, let birth date hierarchy be {D/M/Y, M/Y, Y, 10Y, C/Y/M/O, *}, where 10Y stands for 10-year interval and C/Y/M/O for child, young, middle age and old age. WHD from D/M/Y to Y is WHD(6,4)=(1+1)/5=0.4. In gender hierarchy, {M/F, *}, WHD from M/F to * is WHD(2,1)=1/1=1. This means that the distortion caused by the generalization of five cells from D/M/Y to Y is equivalent to the distortion caused by the generalization of two cells from M/F to *.

(2). Height Weight: $w_{j,j-1} = \frac{1}{(j-1)^{\beta}}$ $(2 \leq j \leq h, \beta \geq 1)$.

For a fixed β, the intuition of this scheme is that the generalization near to the top should give greater distortion compared with the generalization far from the top. Thus, we formulate the height weight scheme, where the weight near to the top is larger and the weight far from the top is smaller. For example, consider a hierarchy: {D/M/Y, M/Y, Y, 10Y, C/Y/M/O, *} for birth date. Let $\beta = 1$. WHD from D/M/Y to M/Y is WHD(6,5)=(1/5)/(1/5 + 1/4 + 1/3 + 1/2 + 1) = 0.087. In gender hierarchy {M/F, *}, WHD from M/F to * is WHD(2,1)=1/1=1. The distortion caused by the generalization of one cell from M/F to * in gender attribute is more than the distortion caused by the generalization of 11 cells from D/M/Y to M/Y in birth date attribute.

In some cases, attributes should be generalized only up to a certain degree or not transformed at all. Otherwise, their values become useless for an application domain. Priorities are used to specify the degree of desired anonymisation of attributes. In some applications, exact values for a specific attribute may be favored while the generalization degree of others is negligible. By specifying priorities the user is able to determine the degree of generalization and information loss s/he is willing to cope with. Attributes with lower priorities are generalized first while attributes



Equally Scaled Attribute Priority: $P_{\text{Gender}} = 0$, $P_{textrmPostCode} = 0.5$, $P_{\text{Age}} = 1$

Figure 1: Generalization Hierarchy with Equally Scaled Attribute Priority

with higher priorities are only generalized when no other solution may be found. Priorities have values in the range [0,1]. The most important attribute has the highest priority value and the differences between any two consecutive priorities values are determined by a pre-defined weight $w'_{j,j-1}$.

Definition 4 (Attribute Priority) *A priority $P_j \in [0,1]$ is assigned to each attribute α_j. Suppose attributes $\alpha_1, \alpha_2, \cdots, \alpha_m$ sorted by their priorities, where α_1 is the highest and α_m is the lowest one and let the weight between attribute α_{j-1} and α_j be predefined, denoted by $w'_{j,j-1}$, where $2 \leq j \leq m$. Then, the priority P_i of attribute α_i is defined as:*

$$P_j = \begin{cases} 1 & \text{if } j = 1 \\ 1 - w'_{j,j-1} \cdot \frac{j-1}{m-1} & \text{if } 2 \leq j \leq m \end{cases}$$

We can similarly define $w'_{j,j-1}$ as $w_{j,j-1}$. For the sake of simplicity, in this paper, we discuss the equally scaled attribute priority, i.e., when $w'_{j,j-1} = 1$ ($2 \leq j \leq m$). The following example illustrates the equally scaled priority values: $P_{\text{Gender}} = 0$, $P_{textrmPostCode} = 0.5$, $P_{\text{Age}} = 1$. According to this priority scheme, Gender is generalized first, Postcode next and Age the last. Because the anonymous solution is obtained after the generalization of Gender and Postcode, so no generalization needed for Age. (see Table 1(b)).

Priorities are used to weight the information loss (distortion) quantifiers of generalized attribute values. Hence, in the final generalization solution the information loss for attribute Age should be much smaller than for attribute Gender. In other words, attribute Gender might be transformed to a more general value than attribute Age, which is the case in our example.

In the following, we define distortions (information loss) caused by the generalization of tuples and tables.

Definition 5 (Weighted Tuple Distortions) *Let $t = \{v_1, v_2, \cdots, v_m\}$ be a tuple and $t' = \{v'_1, v'_2, \cdots, v'_m\}$ be a generalized tuple of t. Let $level(v_j)$ be the domain level of v_j in the attribute hierarchy of α_j and P_j is the attribute priority of α_j. Then, the distortion of this generalization is defined as:*

$$Distortion(t, t') = \sum_{j=1}^{m} P_j \cdot WHD(level(v_j), level(v'_j))$$

Different from (Li et al. 2006), our distortion function is the weighted version which specifies the attribute priority. For example, let the weights of WHD be defined by the uniform weight, attribute Gender be in hierarchy of {M/F, * } and attribute Postcode be in hierarchy of {dddd, ddd*, dd**, d***, * }. $P_{\text{Gender}} = 0$, $P_{textrmPostCode} = 0.5$, and $P_{\text{Age}} = 1$ are the equally scaled priority values. Let t_3 be tuple 3 in Table 1(a) and t'_3 be tuple 3 in Table 1(b). For attribute Gender, WHD=1. For attribute Postcode, WHD=1/4=0.25. For attribute Age, WHD=0. Therefore, Distortion(t_3, t'_3)=1*0+0.25*0.5+0*1=0.125. Compare with (Li et al. 2006), our measurement causes less distortion.

Similar with (Li et al. 2006), we can define the total distortion for the table.

Definition 6 *Let T' be generalized from table T, t_j be the j^{th} tuple in T and t'_j be the j^{th} tuple in T'. Then, the distortion of this generalization is defined as:*

$$Distortion(T, T') = \sum_{j=1}^{|T|} Distortion(t_j, t'_j)$$

where $|T|$ is the number of tuples in T.

From Table 1(a) and (b), Distortion(t_1, t'_1) =Distortion(t_2, t'_2)=Distortion(t_5, t'_5)=Distortion(t_6, t'_6)=0, and Distortion(t_3, t'_3)=Distortion(t_4, t'_4)=0.125. So, the total distortion between the two tables is Distortion(T,T')=0.125+0.125=0.25.

Definition 7 *All allowable values of an attribute form a hierarchical value tree. Each value is represented as a node in the tree, and a node has a number of child nodes corresponding to its more specific values. Let t_1 and t_2 be two tuples. t_c is the closest common generalization of t_1 and t_2 for all attributes α_j ($1 \leq j \leq m$). Then, t_c is defined as:*

$$v_c^j = \begin{cases} v_1^i & \text{if } v_1^i = v_2^i \\ \text{the closest common ancestor} & \text{Otherwise} \end{cases}$$

For example, Figure 1 shows a hierarchical structure with four domain levels. Let t_1={male, young, 4351} and t_2={female, young, 4352}, then t_c={*, young, 435*}. Now, we define the distance between two tuples.

Definition 8 *Let t_1, t_2 be two tuples and t_c be their closest common generalization. Then, the distance between t_1 and t_2 is defined as:*

$$Dist(t_1, t_2) = Distortion(t_1, t_c) + Distortion(t_2, t_c)$$

For example, let the weights of WHD be defined by the uniform weight, attribute Gender and Postcode be in hierarchy shown in Figure 1. t_1={male, young, 4351} and t_2={female, young, 4352}. Then, t_c={*, young, 435*} and Dist(t_1,t_2)=Distortion(t_1, t_c)+Distortion(t_2,t_c)=0.125+0.125=0.25. We discuss some properties of the distance metric in the following.

Theorem 1 *The distance between two tuples t_1 and t_2 Dist(t_1, t_2) satisfies the following properties:*

(1)Dist(t_1, t_1)=0;

(2)Dist(t_1, t_2)=Dist(t_2, t_1);

(3)Dist(t_1, t_3)$\leq$Dist(t_1, t_2)+Dist(t_2, t_3)

3 KNN-Clustering Problem

Typical clustering problems require that a specific number of clusters be found in solutions. However, the k-anonymity problem does not have a constraint on the number of clusters; instead, it requires that each cluster contains at least k tuples. Thus, we pose the k-anonymity problem as a clustering problem, referred to as k-Nearest Neighbor(KNN) Clustering Problem.

Definition 9 (KNN Clustering Problem) *The k-Nearest Neighbor(KNN) Clustering Problem is to find a set of clusters from a given set of n tuples such that each cluster contains k ($k \leq n$) data points and that the average intra-cluster distances is minimized.*

The distance functions that measure the similarities among data points and the cost function which the clustering problem tries to minimize are the heart of every clustering problem. The distance functions are usually determined by the type of data being clustered, while the cost function is defined by the specific objective of the clustering problem. In this section, we describe our distance and cost functions which have been specifically tailored for the priority driven k-anonymisation problem.

Distance Function: A distance function in a clustering problem measures how dissimilar two data points are. As the data we consider in the k-anonymity problem are person-specific records that typically consist of both numeric and categorical attributes, we need a distance function that can handle both types of data at the same time. We are aware that the distance metric *Dist*() defined in Section **??** can deal with both categorical and numeric attributes, so we introduce a density metric called k-Nearest Neighbor(KNN) distance which is defined as follow:

Definition 10 (KNN Distance) *Let T be a set of tuples and t be a tuple in T, and $DistK(i)$($i = 1, 2, \cdots, k$) be the minimal k values in all Dist(t, t_j)($1 \leq j \leq |T|$). Then, the KNN distance of t is defined as:*

$$DistKNN(t) = \frac{\sum_{i=1}^{k} DistK(i)}{k}$$

where $|T|$ is the number of tuples in T.

Definition 11 (Density) *Let $DistKNN(t)$ be the KNN distance of tuple $t \in T$. Then, the density of t is defined as:*

$$Density(t) = \frac{1}{DistKNN(t)}$$

The smaller the distances between t and other records around it are, the larger the density of t is. The tuple (record) with larger density will be made as a cluster center with high probability because the cluster has a smaller distortion. Next, we discuss the cost function which the KNN Clustering Problem.

Cost Function: As the ultimate goal of our clustering problem is the k-anonymisation of data, we formulate the cost function as in Definition 6 to represent the amount of distortion (i.e., information loss) caused by the generalization process. Note that in the rest of the paper, for a table T, to make the notions simple, we use Distortion(T) rather than Distortion(T,T') to represent the distortion between T and its generalized form T'.

As in most clustering problems, an exhaustive search for an optimal solution of the KNN-clustering problem is potentially exponential. Since the k-anonymity problem is shown NP-hard (Aggarwal et al. 2005, Meyerson &Williams 2004, Sun et al. 2008b), and it is also a special case of priority driven k-anonymity problem when each attribute has the same priority, so the priority driven k-anonymity problem is NP-hard as well. Because of the hardness of the problem, we propose a simple and efficient density-based clustering algorithm. The idea is as follows. Given a set T of $|T|$ records, the choice of cluster center points can be based on the distribution density of data points. We pick a record $t \in T$ whose density is the maximal and make it as the center of a cluster C. Then we add $k-1$ records which have minimal distance with t to C. Choose the next cluster center and repeat the clustering process until there are less than k records left. We then iterate over these leftover records and insert each record into a cluster with respect to which the increment of the distortion is minimal.

How to choose the next cluster center is another important issue when one iteration has finished, because we consider that the next cluster center is a record which has the maximal density in remainder records. The next cluster center is not in the k-nearest-neighbor records of this center, thus a principle of choosing the next cluster center is needed.

Definition 12 *Let T be a set of records, t_C be a center of cluster C and t'_C be the next cluster center. The choice of $t'_C \in T \setminus C$ must satisfy the follow two requirements:*

$$Density(t'_C) = max\{Density(t_i), t_i \in \{T \setminus C\}\}$$

$$Dist(t_C, t'_C) > DistKNN(t_C) + DisKNN(t'_C)$$

As the algorithm finds a cluster with exactly k records as long as the number of remaining records is equal to or greater than k, every cluster contains at least k records. If there remain less than k records, these leftover records are distributed to the clusters that are already found. That is, in the worst case, $k-1$ remaining records are added to a single cluster which already contains k records. Therefore, the maximum size of a cluster is $2k-1$. The total time complexity is in $O(n^2)$.

The focus of most k-anonymity work is heavily placed on the QID, and therefore other attributes are often ignored. However, these attributes deserve more careful consideration. In fact, we want to minimize the distortion of QID not only because the QID itself is meaningful information, but also because a more accurate QID will lead to good predictive models on the transformed table. In fact, the correlation between the QID and other attributes can be significantly weakened or perturbed due to the ambiguity introduced by the generalization of the QID. Thus, it is critical that the generalization process does preserve the discrimination of classes using QID. Iyengar (Iyengar 2002) proposed the classification metric (CM) as:

$$\text{CM} = \frac{\sum_{\text{all rows}} \text{Penalty(row } r)}{|T|}$$

Attribute	Distinct Values	Generalizations	Height
Age	74	5-,10-20-year range	5
Work class	8	Taxonomy Tree	3
Education	16	Taxonomy Tree	4
Country	41	Taxonomy Tree	3
Marital Status	7	Taxonomy Tree	3
Race	5	Taxonomy Tree	3
Occupation	14	Taxonomy Tree	2
Gender	2	Suppression	21
Salary class	2	Suppression	1

Table 2: Features of Adult Dataset

where $|T|$ is the total number of records, and Penalty(row r)=1 if r is suppressed or the class label of r is different from the class label of the majority in the QI-group.

Inspired by this, the algorithm is now forced to choose clusters with the same class label for a record, and the enforcement is controlled by the row penalty. We show the results in Section 4 that our modified algorithm can effectively reduce the cost of classification metric without increasing much distortion.

4 Empirical Study

The main goal of the experiments was to investigate the performance of our approach in terms of data quality, efficiency. To evaluate our approach, we also compared our implementation with another algorithm, namely the *median partitioning algorithm*(MPA) proposed in (Leferve et al. 2006). We conduct the experiments with two type of distortion measurement discussed in Section **??**-weighted hierarchical distance and attribute priority.

In our experiment, we adopted the publicly available data set, Adult Database, at the UC Irvine Machine Learning Repository[1], which has become the benchmark for evaluating the performance of k-anonymity algorithms adopted by (Leferve et al. 2005, 2006, Fung et al. 2005, Sun et al. 2008c). We eliminated the records with unknown values. The resulting data set contains 45222 tuples. For k-anonymisation, we considered {age, work class, education, marital status, occupation, race, gender, and native country} as the QID. In addition to that, we also retained the salary class attribute to evaluate the classification metric (CM). The feature of QIDs is shown in Table 2.

We report experimental results on the Density-Based Clustering Algorithm(DBCA) and its modification to reduce classification error(DBCA:CM) for data quality and execution efficiency.

Figure 2 reports the Total distortion of the three algorithms (MPA, DBCA, and DBCA:CM). For increasing values of k. As the figure illustrates, the DBCA:CM algorithm results in the least distortion for all k values. Note also that the distortion of DBCA is very close to the modified DBCA:CM. The superiority of our algorithms over the MPA results from the fact that the MPA considers the proximity among the data points only with respect to a single dimension at each partitioning.

Figure 3 shows the experimental result with respect to the CM metric. As expected, the DBCA:CM modified to minimize classification errors outperforms all the other algorithms. Observe that even without the modification, the DBCA still produces less classification errors than the MPA for every k value. We also measured the execution time of the algorithms for different k values. The results are shown in Figure 4. Even though the execution time for the DBCA is higher than the MPA, we believe that it is still acceptable in practice as k-anonymisation is often considered an off-line procedure.

5 Conclusion and Future Work

In this paper, we propose a priority-driven anonymisation technique that allows to specify the degree of acceptable distortion for each attribute separately. We define generalization distances between tuples to characterize distortions by generalizations, which works for both numerical and categorical attributes. Further, we propose a density-based clustering technique to minimize information loss and thus ensure good data quality. We experimentally show that the proposed method is more scalable and causes significantly less distortions than an optimal k-anonymity method.

In the future work, we focus on two important extensions. First, we would try to extend this priority driven anonymisation framework to other privacy requirements, like (p^+, α)-sensitive k-anonymity (Sun et al. 2008c), l-diversity (Machanavajjhala et al. 2006), (α, k)-anonymity (Li et al. 2006) and t-closeness (Li et al. 2007), etc, to make it a systematic approach. Second, we could like to do more experimental studies to compare the performance with other clustering methods (Byun et al. 2006).

Acknowledgement

We would like to thank anonymous reviewers for their useful comments on this paper. This research was supported by Australian Research Council (ARC) grant DP0774450 and DP0663414.

References

Aggarwal, G & Feder, T & Kenthapadi, K & Motwani, R & Panigrahy, R & Thomas, D & Zhu, A (2005), Anonymizing tables, *in* In Proc. of the 10th International Conference on Database Theory, pp. 246-258, Edinburgh, Scotland.

Aggarwal, G & Feder, T & Kenthapadi, K & Zhu, A & Panigrahy, R &Tomas, D (2006), Achieving anonymity via clustering in a metric space, *in* in Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, 2006.

Aggarwal, C. C (2005), On k-anonymity and the curse of dimensionality, Proceedings of the 31st international conference on Very large data bases, pages 901-909. VLDB Endowment, 2005.

Agrawal, D & Aggarwal, C. C (2001), On the design and quantification of privacy preserving data mining algorithms, in Proceedings of the twentieth ACM SIGMOD-SIGACTSIGART symposium

[1] available at `www.ics.uci.edu/~mlearn/MLRepository.html`

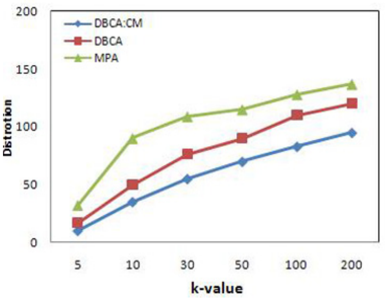


Figure 2: Distortion Metric

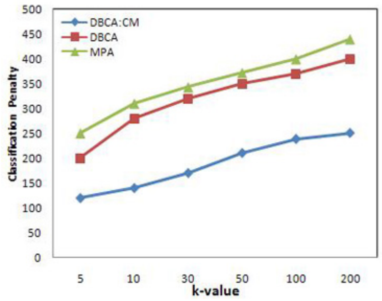


Figure 3: Classification Metric

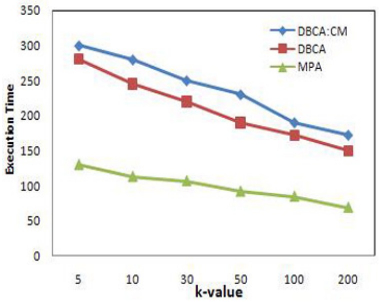


Figure 4: Running Time

on Principles of database systems, pages 247-255, New York, NY, USA, 2001. ACM Press.

Byun, J. W., & Kamra, A & Bertino, E & Li, N (2006), Efficient k-Anonymity using Clustering Technique, CERIAS Tech Report 2006–10, 2006.

Fung, B & Wang, K & Yu, P (2005), Top-down specialization for information and privacy preservation, In Proc. of the 21st International Conference on Data Engineering, Tokyo, Japan.

Bayardo, R & Agrawal, R (2005), Data privacy through optimal k-anonymity, In Proceedings of the 21st International Conference on Data Engineering (ICDE) 2005.

Domingo-Ferrer, J & Torra, V (2005), Ordinal, continuous and heterogeneous k-anonymity through microaggregation, Data Mining and Knowledge Discovery, 11(2):195-212, 2005.

Domingo-Ferrer, J & Mateo-Sanz, J. M (2005), Practical data-oriented microaggregation for statistical disclosure control, IEEE Transactions on Knowledge and Data Engineering, 14(1):189-201, 2002.

Iyengar (2002), Transforming data to satisfy privacy constraints. In ACM Conference on Knowledge Discovery and Data mining, 2002.

Leferve, K & Dewitt, D & Ramakrishnan, R. (2005), Incognito: Efficient Full-Domain k-Anonymity, ACM SIGMOD International Conference on Management of Data, June 2005.

Leferve, K & Dewitt, D & Ramakrishnan, R. (2006), Mondrian multidimensional k-anonymity. In International Conference on Data Engineering, 2006.

Machanavajjhala, A & Gehrke, J & Kifer, D & Venkitasubramaniam, M (2006), l-Diversity: Privacy beyond k-anonymity, In ICDE, 2006.

Meyerson, A &Williams, R (2004), On the complexity of optimal k-anonymity, in Proc. of the 23rd ACM-SIGMOD-SIGACT-SIGART Symposium on the Principles of Database Systems, pp. 223-228, Paris, France, 2004.

Li, N & Li, T & Venkatasubramanian, S (2007), t-Closeness: Privacy Beyond k-Anonymity and l-Diversity, In 23rd IEEE International Conference on Data Engineering (ICDE), April 2007

Samarati, P (2001), Protecting respondents' identities in microdata release, IEEE Transactions on Knowledge and Data Engineering, 13(6):1010-1027. 2001

Sun, X & Li, M & Wang, H & Plank, A (2008), An efficient hash-based algorithm for minimal k-anonymity. In: 31st Australasian Computer Science Conference (ACSC 2008), 22-25 Jan 2008, Wollongong, NSW, Australia.

Sun, X & Wang, H & Li, J (2008), On the complexity of restricted k-anonymity problem, The 10th Asia Pacific Web Conference(APWEB2008), LNCS 4976, pp: 287-296, Shenyang, China. 2008.

Sun, X & Wang, H & Li, J & Traian, T. M & Ping, L (2008), (p^+, α)-sensitive k-anonymity: a new enhanced privacy protection model. In 8th IEEE International Conference on Computer and Information Technology (IEEE-CIT 2008), 8-11 July 2008, Sydney, Australia. pp:59-64.

Sweeney, L (2002), Achieving k-anonymity Privacy Protection Using Generalization and Suppression, International Journal of Uncertainty, Fuzziness and Knowledge-Based System, 10(5) pp. 571-588, 2002.

Sweeney, L (2002), k-anonymity: A Model for Protecting Privacy, International Journal on Uncertainty Fuzziness Knowledge-based Systems, 10(5), pp 557-570, 2002.

Sweeney, L (2000), Uniqueness of simple demographics in the u.s. population, Technical report, Carnegie Mellon University, 2000.

Wong, R & Li, J & Fu, A & Wang, K (2006), (α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing, *KDD 2006*: 754-759.

Li, J & Wong, R & Fu, A & Pei, J (2006), Achieving k-Anonymity by clustering in attribute hierarchical structures. In: 8th International Conference on Data Warehousing and Knowledge Discovery, 4-8 Sept 2006, Krakow, Poland.