K-anonymity Through the Enhanced Clustering Method

Md Ileas Pramanik
Department of Information Systems
City University of Hong Kong
Hong Kong
mpramanik2-c@my.cityu.edu.hk

Raymond Y.K.Lau
Department of Information Systems
City University of Hong Kong
Hong Kong
raylau@cityu.edu.hk

Wenping Zhang
Department of Information Systems
City University of Hong Kong
Hong Kong
wzhang23-c@my.cityu.edu.hk

Abstract—**With the rise of the Social Web, there is increasingly more tendency to share personal records, and even make them publicly available on the Internet. However, such a wide spread disclosure of personal data has raised serious privacy concerns. If the released dataset is not properly anonymized, individual privacy will be at great risk. K-anonymity is a popular and practical approach to anonymize datasets. In this study, we use a new clustering approach to achieve k-anonymity through enhanced data distortion that assures minimal information loss. During a clustering process, we include an additional constraint, minimal information loss, which is not incorporated into traditional clustering approaches. Our proposed algorithm supports a data release process such that data will not be distorted more than they are needed to achieve k-anonymity. We also develop more appropriate metrics for measuring the quality of generalization. The new metrics are suitable for both numeric and categorical attributes. Our experimental results show that the proposed algorithm causes significantly less information loss than existing clustering algorithms.**

Keywords-Privacy; K-anonymity; Clustering; Generalization; Suppression;

I. Introduction

Organizations, such as hospitals, have been generating a vast amount of operational data and information, where most of the generated data are useful only when they are shared and analyzed with other related datasets. However, this type of data normally contains individual details and personal information that may be revealed in sharing and analyzing processes. Conventionally, in order to address the privacy concern, identifying attributes are excluded from the released datasets. Recent research [1][2] has demonstrated that such protections are inadequate due to the existence of quasi-identifiers (e.g., Sex, Age, Date of birth) in the released dataset. The quasi-identifiers (QID) are the set of attributes that can be combined with data from other sources to identify personal records [2].

To address this threat, cryptographic approach is an option because this technique is able to hide data from unauthorized access. Cryptographic techniques generally change the content of records too much to restrict data access [3]. Therefore, data utility is severely affected by this method. However, different cryptographic data privacy methods (e.g., [4][5]) tailor some particular data mining tasks through compromising data utility.

Another privacy protection approach is the perturbation method, which is suitable for numeric attributes [6][7][8]. When the attributes are categorical then such approaches are not adequate to protect privacy effectively [9]. Recently, a new method of protecting data privacy in relation to both categorical and numerical attributes called k-anonymity [10] has gained more popularity. In the k-anonymity method, quasi-identifiers that leak private information are suppressed and generalized so that every record in the released data is identical to at least other k-1 records with respect to quasi-identifiers [11]. Therefore, most existing k-anonymity methods use generalization and suppression for preserving privacy in the released datasets. The k-anonymity is a simple and practical approach and so it attracts a number of researchers to do more work and design a number of algorithms [12][25] using this method. K-anonymity mainly uses generalization (conversion of specific data into a range) and suppression (removal of data from the original dataset) that inadvertently lead to loss of data utility. Data utility and data privacy conflict with each other. Hence, a proper tradeoff between privacy and data utility emerges.

The objective of this paper is to propose a novel clustering approach to achieve k-anonymity with minimum information loss, where no data records are completely suppressed. We mainly exclude suppression in our proposed model because the suppression seriously damages the data quality and utility as well. At the time of data clustering, most existing methods exclude data records from the released dataset to achieve anonymity, whereas there is no need to remove any complete record from the released data in our proposed method. Intuitively, we can visualize the comparison between the existing and proposed methods as in Fig. 1, where records in the inter cell gap (in existing methods) are removed in the released microdata. On the other hand, there are no inter cell gaps in the proposed approach and so there is no need to remove any data record. In this study, we developed an algorithm for this purpose and demonstrate that the proposed method provides k-anonymity with minimal data distortion. Moreover, to measure the information loss we developed more appropriate metrics to measure data distortion accurately.

The remainder of this paper is organized as follows: In section 2, we present some preliminary definitions. In section 3, we introduce metrics for measuring the quality of k-anonymization. In section 4, we illustrate an enhanced k-anonymity algorithm. Section 5 presents empirical study, and finally section 6 concludes the paper.

978-1-5090-6119-8/16 $31.00 © 2016 IEEE
DOI 10.1109/ICEBE.2016.11

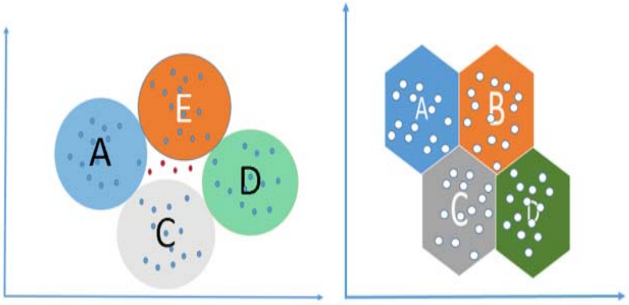
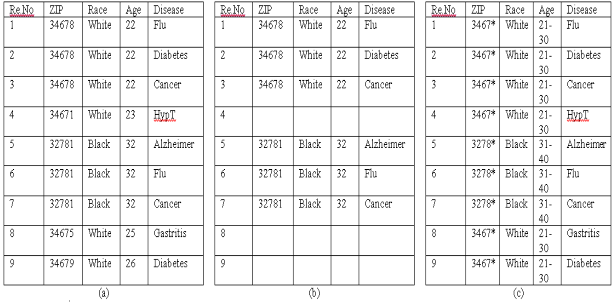


Fig. 1 Left: Existing k-anonymity methods with suppression (the red four data points are suppressed); Right: our proposed k-anonymity method without suppression

II. Preliminary Definitions

The process of k-anonymity is to remove all identifying attributes from the released data and then generalize/suppress the QIDs. In the generalization approach, multiple values are combined to a single generalized value. The number of distinct tuples are decreased in generalization, thereby the size of cluster is increased with the same values [13]. The generalization method modifies the dataset such that the total number of tuples remains unchanged and all values of an attribute belong to the same domain [14]. Suppression is another complementary approach to provide k-anonymity, where data records are removed from released datasets. These two methods are widely used in the context of statistical databases as well [15][16]. Though suppression is more useful to achieve anonymity, it distorts data more severely than generalization approaches.

Fig. 2(a) shows a private table (PT) with nine records, three quasi-identifier attributes (ZIP, Race, Age) and one sensitive attribute (Disease). We can achieve 3-anonymity from PT by using either suppression as in Fig. 2(b) or generalization as in Fig. 2(c). When comparing the row data with released data, we can see the number of tuples is the same in the case of generalization, whereas in suppression three records are missing. We use generalization in our proposed method because generalization has the advantage of allowing the release of all single records in the released dataset. On the other hand, suppression removes records from the released table, which is the main cause of increased information loss and decreased data utility.



Re.No	ZIP	Race	Age	Disease
1	34678	White	22	Flu
2	34678	White	22	Diabetes
3	34678	White	22	Cancer
4	34671	White	23	HypT
5	32781	Black	32	Alzheimer
6	32781	Black	32	Flu
7	32781	Black	32	Cancer
8	34675	White	25	Gastritis
9	34679	White	26	Diabetes

(a)

Re.No	ZIP	Race	Age	Disease
1	34678	White	22	Flu
2	34678	White	22	Diabetes
3	34678	White	22	Cancer
4				
5	32781	Black	32	Alzheimer
6	32781	Black	32	Flu
7	32781	Black	32	Cancer
8				
9				

(b)

Re.No	ZIP	Race	Age	Disease
1	3467*	White	21-30	Flu
2	3467*	White	21-30	Diabetes
3	3467*	White	21-30	Cancer
4	3467*	White	21-30	HypT
5	3278*	Black	31-40	Alzheimer
6	3278*	Black	31-40	Flu
7	3278*	Black	31-40	Cancer
8	3467*	White	21-30	Gastritis
9	3467*	White	21-30	Diabetes

(c)

Fig. 2 (a). Private table (PT); (b) Record suppression of PT; (c) Attribute generalization of PT

Definition 1 (Quasi-Identifier). Let, T be a table with attributes $\{A_1, A_2, \ldots\ldots A_n\}$and records $\{t_1, t_2, \ldots t_m\}$, where each record corresponds to an individual data. A quasi-identifier of table T is a minimum set of attributes $\{A_i \ldots\ldots A_j\} \subseteq \{A_1, A_2, \ldots\ldots A_n\}$ such that $\cup_i^j A_x$=$\exists_t(T)$ [2].

For example, Attribute set {ZIP, Race, Age} in private table (PT) is a quasi-identifier. When the relevant values from these attributes are joined, it can reveal at least one individual. Suppose for the ZIP, Race, and Age values are 34675, white, and 25 respectively, then they actually reveal record no 8.

Definition 2 (Equivalence Class). An equivalence class of a table (T) is the set of records that have identical value of quasi-identifier attributes.

For example, first three records in Fig. 2a are equivalence with respect to quasi-identifiers, IDs {ZIP,Race,Age}.

Definition 3 (*K*-anonymity). Let, a table T with quasi-identifier attributes $\{A_i, \ldots\ldots A_j\}$ and records $\{t_1, t_2, \ldots t_m\}$. T is to satisfy *k*-anonymity iff at least k number of records are identical with respect to QI ($t_1 = t_2 = \cdots .. = t_k\ w.r.t\ all\ QI$) [14].

For example, Fig. 2c is a 3-anonymity view of private table PT, because minimum size of equivalence class is not less than 3. Therefore, none can identify any distinct record with a probability greater than 1/3 from Fig. 2c.

In the data table, each attribute contains a set of values. Generalized attributes means mapping the attribute values which are stated by the means of a generalization relationship ($\leq_g$). Let two values V_i and V_j relationship $V_i \leq_g V_j$ describes the fact that the values V_i are generalized by V_j. The value of V_j generalizes V_i values, when it satisfy the following conditions (Samarati, P.2001).

1. $(\forall V_i, V_i \in V_j)\vee(\exists V_i, V_j \notin V_i)$
2. All maximal values of Attribute are singleton.

For example, in Fig. 2c, age values 22,23,25,26 are generalized by the range [21-30], and 32,34,38 are generalized by the range value [31-40].

By the same token, an attribute A_i is generalized by A_j, ($A_i \leq_g A_j$) iff $\exists V_i, V_j\ \ V_i \leq_g V_j$ where $V_i \in A_i$ and $V_j \in A_j$.

Definition 4 (Generalized Table). Let, $T_i\ and\ T_j$ be two tables with n number of attributes $\{A_1, A_2, \ldots\ldots A_n\}$ and m number of tuples $\{t_1, t_2, \ldots\ldots t_m\}$. Table T_i is generalized by table T_j $(T_i \leq_g T_j)$ iff

1. $\exists\ V, V_z\ (A_{i,}) \leq_g V_z(A_j)$ where $V(A)$ denotes the value V of attribute A.
2. $\exists\ A, A_z(t_i) \leq_g A_z(t_j)$ where $A(T)$ indicates the attribute A of tuple t.

3. $\exists\, t, t_z(T_i) \leq_g t_z(T_j)$ where $t(T)$ indicates the tuple t of table T.

According to above definition, Fig. 2c satisfies all above conditions of generalized table. Therefore, that table generalizes table in Fig. 2a. For assuring data utility, minimum information loss is required in *k*-anonymization that is called enhanced *k*-anonymity. In our method, *k*-minimal generalization approach [13] is used in original data to get enhanced *k*-anonymization. *K*-minimal generalization approach can be defined as follows:

Definition 5 (*K*-minimal generalization). Let, T_i and T_j be two tables with n number of attributes $\{A_1, A_2, \ldots \ldots A_n\}$ and m number of records $\{t_1, t_2, \ldots \ldots t_m\}$. Table T_j is said to be a *k*-minimal generalization of T_i, iff

1. T_j satisfies *k*-anonymity (According to definition 3)
2. $\forall\, T_{z(z \neq j)} : T_i \leq_g T_z,\ where\ T_z$ does not satisfy k anonymity, unless $T_j \leq_g T_z$.

The main objective of *k*-minimal generalization is to achieve *k*-anonymity through minimal change of attribute values. Moreover, a table T_i is minimal generalization of itself if it satisfies *k*-anonymity for all k [14].

III. Distance and Information Loss Metrics

A number of information loss metrics exist in the current literature. The Discernibility Metric (DM) [17] mainly measures the cardinality of the clustered data. Although clusters with few records are desirable, DM does not consider the distance of records in the quasi-identifier space. The Generalized Loss Metric [18] and the similar Normalized Certainty Penalty (NCP) [19] are the more popular matrices for measuring the quality of anonymization. In this article, we do not use them due to their higher cost. In NCP, the cost of finding information loss between two databases is comparatively high [19]. Classification metric (CM) is also another quality measurement technique introduced by Iyneger [17] to optimize a *k*-anonymous dataset for training a classifier. CM measures the information loss by adding the individual penalties for each tuple over the total number of records. We don't use CM because it is not clear to us how we can extend CM to support general purpose applications. Another loss measurement method was proposed by Aristides and Tamir in [20], who called their method entropy measurement. Their method is a theoretical measurement method, which is difficult to implement in real-world datasets. However, for measuring distortion between original and changed data, we use weighted hierarchical distance [9], which is a simpler and more pragmatic approach. We also include weights for different attributes that are calculated from the generalization level and number of attributes in the experimental datasets. The weight distance between two categorical values is as follows:

Definition 6 (Weighted distance between categorical values). Let A be a categorical attribute and h is the height of weight hierarchy of A. Here, $w_{i,i+1}$ is the weight from level i to level $i+1$ where level, $i = 0,1,2,3,4, \ldots \ldots, h$. If v is the value of attribute A, then the Weighted Hierarchical Distance (WHD) between the value of original data (v_o) and the value of anonymized data (v_a) is defined as:

$$WHD(v_0, v_a) = \frac{\sum_{i=a+1}^{o} w_{i,i+1}}{\sum_{i=0}^{h-1} w_{i,i+1}}$$

Height and weight levels of a sample hierarchy are shown in Fig 3. According to [9], there is two simple schemes to measure weights. One is uniform weight where all weights are equal to 1 ($w_{i,i+1} = 1, \forall\, i$). If ZIP hierarchy is {3421,342*, 34**, 3***, *} corresponds to {suburb, city, region, state, unknown}, respectively, WHD from suburban to state is (1+1+1)/4=0.75. The uniform weight scheme is quite simple, which captures uniform weight levels, though the distortion at different levels are not similar. In this scheme, data distortion for the generalization near to the root is equal to the generalization far from the root. Therefore, the accuracy of this scheme is not quite good in some contexts. This gap mainly motivated to use another scheme of weight, called Height Weight. In Height Weight scheme, generalization near to the root presents more distortion than generalization far from the root. Therefore, the Height Weight scheme can be defined as:

$$w_{i,i+1} = \frac{1}{h} \sum \frac{1}{(i+1)^{\gamma}}$$

$$where, i = \{0,1,2,3 \ldots \ldots h-1\}$$

where γ defines the closeness of generalized value with root and leaf nodes. With fixed $\gamma = 1$, for the case of Height Weight, in above example, WHD from suburban to state is (1/5+1/4+1/3)/4=0.196. When generalized value is more near to the leaf nodes then γ will be higher. Similarly, the value of γ will be close to 1 if generalized value is far from the leaf nodes, but near to the root node.

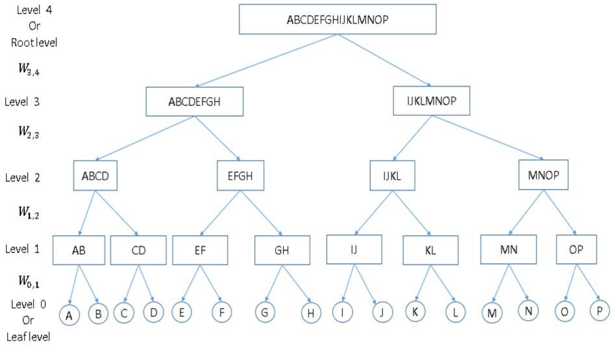


Fig. 3 Hierarchy with height and weight levels

Definition 7 (Weighted distance between Numeric values). Let A be a numeric attribute with finite numeric attribute domain, N. Like [21], if v is the value of attribute

A, then the distance between two different records' values v_o and v_a is defined as:

$$\text{dist}(v_o, v_a) = \frac{|v_o - v_a|}{|N|}$$

Where, $|N| = |v_{max} - v_{min}|$. The age is a numeric attribute in table-2a. The distance between records 7 and 8 in terms of age is 7/10=0.7.

Definition 8 (Distance between two records). Let, a record t_1 with values $v_i, c_i \ where\ i = 1,2,3, \ldots., N$ is generalized by record t_1' with values v_i', c_i', where v and c are numerical and categorical values respectively. The distance between two records is defined as:

$$Dist(t_1, t_1') = \sum_{i=1}^{N} WHD(c_i, c_i') + \sum_{i=1}^{N} dist(v_i, v_i') \qquad (1)$$

Importantly, when any record is suppressed from the original table, the distance between original and suppressed records will be maximum in every attributes, and for this reason, we avoid record suppression in our proposed method.

Moreover, we include different weights for different quasi-identifier (QID) attributes, because the distortion caused by the generalization are not same for all attributes, through the level of generalizations are same. For example, generalization of date of birth (DOB) from (DD/MM/YY) to MM/YY is different from generalization of gender from (M/F) to * in terms of distortion effect, though both cases have done single level generalization. The former case still keeps significant information for DOB whereas, the latter losses total information about gender. Therefore, different QIDs should contain different weight, which are extracted from total generalization level and the number of QIDs in the experimental dataset.

Definition 9 (QID Weight). Let a table T with m number of QIDs $\{A_1, \ldots \ldots . A_m\}$. If the total level of generalization for different attributes are represented by $\{L_{A_1}, \ldots \ldots . L_{A_m}\}$ then the QID weights(WID) are defined as-

$$WID(A_i) = \left(1 - \frac{(L_{A_i})^m}{\sum_{i=1}^{m} (L_{A_i})^m}\right) \qquad (2)$$

For example, a table has three QIDs (Race, ZIP, BOD), and their individual generalization steps are Race {W/B, *}, ZIP{NNNN, NNN*, NN**, N***, ****}, and BOD{D/M/Y,M/Y,Y,10Y}. For any QID, the total level of generalization is the required level of generalization to reach from the leaf to the root level. Therefore, total level generalization for Race is 1, for ZIP is 4 and for BOD is 3. By using (2), we can define the weight for Race attribute, ZIP attribute, and DOB attribute as follows:

$$WID(Race) = (1 - 1^3/(1^3 + 4^3 + 3^3)) = 0.98, WID(ZIP) = (1 - 4^3/(1^3 + 4^3 + 3^3)) = 0.3, and, WID(DOB) = (1 - 3^3/(1^3 + 4^3 + 3^3)) = 0.71$$

IV. Enhanced k-anonymity Algorithm

In this section, we propose an enhanced clustering solution (KOC) for achieving *k*-anonymity in multi-dimensional contexts. An algorithm for enhanced *k*-anonymity is shown in Fig. 5. Before clustering the datasets, we have to find the exact centre of a cluster. We measure the cluster-Centre, where centres are measured by n-closeness as follows.

Closeness distance for a record (t_i) is defined as-

$$DistC(t_i) = \frac{\sum_{j=1}^{N-1} (t_i, t_j)}{N - 1} \quad i \neq j.$$

So, the n-closeness of t_i is defined as-

$$Ncloseness(t_i) = \frac{1}{DistC(t_i)} \qquad (3)$$

Fig. 4 shows the algorithm for extracting cluster-centres. By using (3), we can find the n closeness value for each record. Then records are sorted with n-closeness values. Records with the most n-closeness values are assigned as cluster-centers in different clusters.

```
Input: N-records $\{t_1, t_2, \ldots . . t_N\}$, Number of clusters $(N_G)$
Output: Cluster centers, $GC = \{GC_1, GC_2, \ldots . . GC_{N_G}\}$
$GC = \emptyset$
for i=1 to n do
    $RC_i = Ncloseness(t_i)$
end for
// $RC_i$ is a closeness vector for each records $(t_i)$.
$X_i^* = Short(RC_i)$
// $X_i^*$ is a sorted vector of $RC_i$ in descending order.
for i= 1 to $N_G$ do
    $GC_i = X_i^*$
    $GC = GC \cup GC_i$
end for
$N = N - N_G$
// cluster centers are excluded form total records
return $GC$.
```

Fig. 4. Pseudocode for extracting cluster center.

```
Input: Original data records $\{t_1, t_2, \ldots \ldots . t_N\}$, Cluster centers, $GC = \{GC_1, GC_2, \ldots . . GC_{N_G}\}$
Output: K-anonymous data, $G' = \{G_1', G_2', \ldots \ldots \ldots G_{N_G}'\}$.
i=$N_G$
$G = \emptyset$
$G' = \emptyset$
repeat
```

```
5.$G_i = forming\_Cluster(GC_i)$
6.$G = G \cup G_i$
7.$G_i' \leftarrow G_i$
8.$G' = G' \cup G_i'$   // $X'$, represent the generalization of X for k-anonymization
9.$i = i - 1$;
10.Until $i > 0$
11.return, $G' = \{G_1', G_2', \ldots \ldots \ldots G_{N_G}'\}$. // k-anonymous data
12.function forming_Cluster ($GC_i$)
13.$GC_i' \leftarrow GC_i$
14.$G_i = \emptyset$
15.repeat
16.$r_j^* = dist_{min}(GC_i', r_j')$
//$r_j^*$ is the record within N unassigned records that produce minimal distortion when it add to //cluster $G_i$
17.$G_i = G_i \cup r_j$
18.N=N-1;
19.until ($|G_i|$=k)  // $|X|$ means the number of records in X.
20. if ($G_i = G_{N_G}$) then
 21. if $|G_i| < k$ then
22.    disperse_records( $\sum_{x=1}^{m} r_x$ )   where m<k.
23.   end if
24.end if
25. return $G_i$
26. end forming_Cluster.
27. function disperse_records ($r_1, r_2, \ldots \ldots r_m$)
28. $i = m$
29. repeat
30. for $j = 1$ to $(N_G - 1)$ do  // excluding last cluster center
31. $r_i^* = dist_{min}(GC_j, r_i')$
//$r_i^*$ is the record within m (m<k) assigned records for last cluster
// records in $G_{N_G}$ are dispersed to other clusters w.r.t. minimal distortion
32. end for
33. $G_j = G_j \cup r_i$
34. m=m-1
35.until m>0
36. end disperse_records,
```

Fig. 5. Algorithm for Enhanced *k*-anonymity

Number of clusters (N_G) is also an important parameter to minimize data distortion. When datasets are partitioned into a large number of clusters, then total distortion for anonymization will reduce. In our distortion metric, when a large cluster is divided into some small clusters, then summation of individual distortion in small clusters must be smaller than the distortion of large clusters. Given a cluster G_1 is split into two clusters $G_{1A}, and\ G_{1B}$, if the distortion of these clusters are $Dist(G_1), Dist(G_{1A})$, and $Dist(G_{1B})$, then $Dist(G_1) > Dist(G_{1A}) + Dist(G_{1B})$. Distortions are reduced because data points are placed in closest clusters when larger clusters are fragmented into small clusters. Data points of cluster G_1 are distributed among small clusters $G_{1A}, and\ G_{1B}$ without missing any data point. Since unclustered records are disperse to the closest cluster through the function, *disperse* at line 27, the total number of records remains the same. Thus, $N(G_1) = N(G_{1A}) + N(G_{1B})$. Moreover, in *k*-anonymization, cluster numbers of data points are at least k but not more that 2*k*-1. If the number of data points is less than *k* then no cluster can satisfy *k*-anonymity and so disperse among other clusters where distortions are minimal. Again, when data points are more than 2*k*-1, then clusters will be further fragmented into two clusters.

In lines 5-6 of the enhanced algorithm, we form clusters where near data points are placed in the same cluster. We exclude the process of cluster-centre extraction from our enhanced *k*-anonymity algorithm. In Fig. 4, we compute the centres of all clusters based on the data closeness. The closeness of records is calculated by using (3). A record with maximum closeness in a cluster will be the centre. Therefore, a central record is closer to all other records in their respective cluster. When we can select cluster-centres more accurately, then generalization loss will be comparatively minimal. At the time of cluster formation, data records include those clusters where data distortion is minimal (lines 16-17). Usually, every cluster contains *k* records (lines 16-19), but sometimes the last cluster may contain less than *k* records. If there are less than *k* records in the last cluster, then records are dispersed from the last cluster to other close clusters in terms of minimal distortion (lines 22,31-33). Finally, clusters may contain more than *k* records but not more than (2*k*-1). Then clusters are individually anonymized in line 7. Lastly, line 11 returns a set that contains anonymized clustering data. The complexity for extracting cluster-centres is $O(NlogN)$. All records are sorted and only $O(N)$ passes are required in line 4. Since, the number of clusters, N_G, is evaluated, each iteration requires $O(1)$ time. After finding the minimum distortion (line 16), time complexity is $O\big(n + n(n+1)\big) \approx O(n^2)$. For dispersing data records, time complexity is $O(n^2)$. Thus, the time complexity for the enhanced *k*-anonymity is $O(n^2)$. In [18] they proved that *k*-anonymization for microdata is NP hard, and our proposed enhanced *k*-anonymity problem is similar.

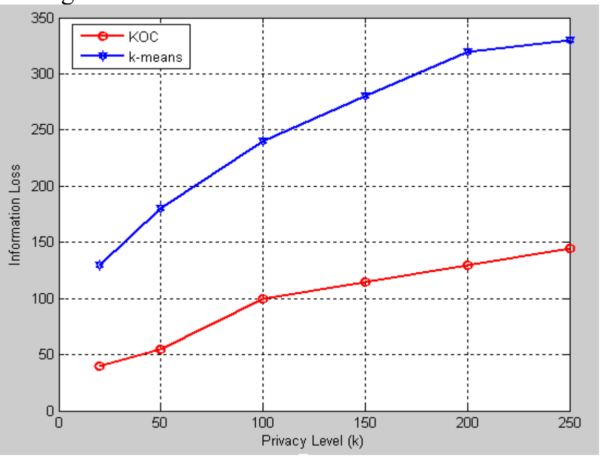
V. Experimental Results

A Pentium IV 2.8 GHz PC with 12GB RAM was used to conduct our empirical study. In this section, we assess our enhanced approaches compared to the existing method where we implement our algorithm in C/C++. In our lab experiments, we tailored the publicly available adult dataset from the UCIrvine Machine Learning repository [22]. Before the experiment, we configured our data similar to [23][12]. Data records with incomplete values were discarded because of limitations in our prototype system. The resulting dataset contained 45,345 records. The schema is summarized in Table 1. There are nine attributes in the dataset where 7 represents the QI and the last two (i.e., Occupation, Salary) are the SAs. In the QI attributes, two (i.e., Age, Education Level) of them are treated as numeric, and others are treated as categorical attributes.

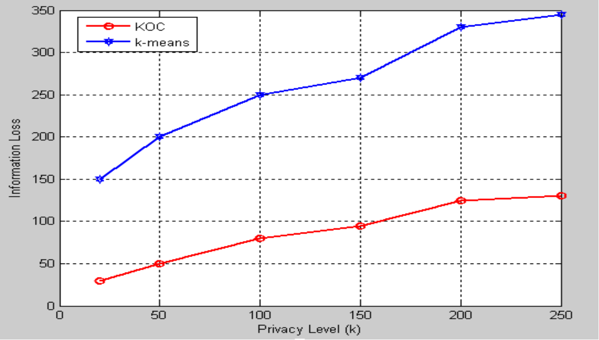
TABLE I. CHARACTERISTICS OF THE ADULT DATASET

	Attribute	**Cardin-ality**	**Type**	**Generalization**	**Generalization level**
1	Age	76	N	Range	3
2	Sex	2	C	Hierarchy	1
3	Education Level	16	N	Range	4
4	Native Country	81	C	Hierarchy	3
5	Work class	8	C	Hierarchy	2
6	Salary	35	C	Hierarchy	1
7	Marital Status	6	C	Hierarchy	1
8	Race	6	C	Hierarchy	1
9	Profession	34	C	Hierarchy	2

We evaluated our algorithm in terms of two measurements: data distortion or information loss and execution time. We also compared KOC algorithm with the best-known clustering algorithm k-means [24] which includes only one condition and contains at least k records, for clustering.



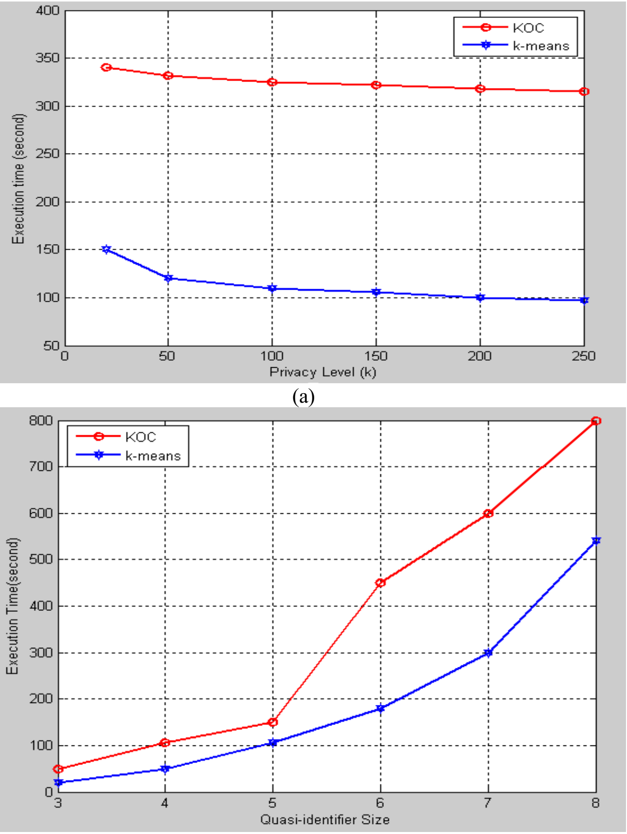
(a)



(b)

Fig. 6 Information Loss with privacy level a) Excluding WID b) Including WID

Fig. 6(a) shows the results with information loss measurement where all attributes of the data table have uniform weight. This experiment result reports that information loss in the KOC algorithm is 2.50 times lower than that in the k-means algorithm for a level of privacy from 20 to 250 on average.



(a)

(b)

Fig. 7. Execution time with a)Privacy Level b) Quasi-identifier Size

Fig. 7(a) shows that the execution time of both algorithms decreases with increasing privacy level (k). Moreover, as shown in Fig. 7(b), the execution time of both algorithms increases with the quasi-identifier size. In both results, execution time of the KOC algorithm is larger than that of the k-means algorithm. However, the time complexity of both algorithms is similar in nature. The execution time of the KOC algorithm is acceptable in those cases where information loss is a fact of major concern. In essence, the experiment shows that the KOC algorithm has better performance in relation to information loss and is acceptable regarding execution time. Therefore, it is experimentally verified that the proposed KOC algorithm can achieve better k-anonymization.

VI. CONCLUSIONS

In this paper, we introduce a novel clustering approach to achieve k-anonymity in terms of minimum information loss. We mainly avoid record suppression in our proposed model because the suppression seriously damages the data quality and utility as well. We also define two general metrics, one excluding and the other including WID, to measure the quality of anonymization.

We experimentally verify that our proposed algorithm causes minimum loss in generalization and less than the k-means clustering algorithm. From the experimental result, we also assure that the measurement of information loss is more accurate when different weights are included in quasi-identifiers. Moreover, we compare these two algorithms in terms of information loss and execution time. The execution time of our proposed algorithm is acceptable in most cases. The information loss of our algorithm is at least 2.50 times smaller than for the k-means algorithm on average. Though the execution time of the KOC algorithm is at a satisfactory level, it is not fully optimized and this will be our extended research of this study in the future.

ACKNOWLEDGEMENT

This work was partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project: CityU 11502115), and the Shenzhen Municipal Science and Technology R&D Funding Scheme – Basic Research Program (Project No. JCYJ20140419115614350).

REFERENCES

[1] Froomkin, A. Michael. "The death of privacy?." Stanford Law Review (2000): 1461-1543.

[2] Sweeney, Latanya. "k-anonymity: A model for protecting privacy." International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10, no. 05 (2002): 557-570.

[3] Lindell, Yehuda, and Benny Pinkas. "Privacy preserving data mining." Journal of cryptology 15, no. 3 (2002): 177-206.

[4] Vaidya, Jaideep, and Chris Clifton. "Privacy-preserving k-means clustering over vertically partitioned data." In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 206-215. ACM, 2003.

[5] Wright, Rebecca, and Zhiqiang Yang. "Privacy-preserving Bayesian network structure computation on distributed heterogeneous data." In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 713-718. ACM, 2004.

[6] Agrawal, Rakesh, and Ramakrishnan Srikant. "Privacy-preserving data mining." In ACM Sigmod Record, vol. 29, no. 2, pp. 439-450. ACM, 2000.

[7] Agrawal, Dakshi, and Charu C. Aggarwal. "On the design and quantification of privacy preserving data mining algorithms." In Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pp. 247-255. ACM, 2001.

[8] Rizvi, Shariq J., and Jayant R. Haritsa. "Maintaining data privacy in association rule mining." In Proceedings of the 28th international conference on Very Large Data Bases, pp. 682-693. VLDB Endowment, 2002.

[9] Li, Jiuyong, Raymond Chi-Wing Wong, Ada Wai-Chee Fu, and Jian Pei. "Achieving k-anonymity by clustering in attribute hierarchical structures." Springer Berlin Heidelberg, 2006.

[10] Wang, Ke, Philip S. Yu, and Sourav Chakraborty. "Bottom-up generalization: A data mining solution to privacy protection." In Data Mining, 2004. ICDM'04. Fourth IEEE International Conference on, pp. 249-256. IEEE, 2004.

[11] LeFevre, Kristen, David J. DeWitt, and Raghu Ramakrishnan. "Incognito: Efficient full-domain k-anonymity." In Proceedings of the 2005 ACM SIGMOD international conference on Management of data, pp. 49-60. ACM, 2005.

[12] Fung, Benjamin, Ke Wang, and Philip S. Yu. "Top-down specialization for information and privacy preservation." In Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on, pp. 205-216. IEEE, 2005.

[13] Samarati, Pierangela. "Protecting respondents identities in microdata release." Knowledge and Data Engineering, IEEE Transactions on 13, no. 6 (2001): 1010-1027.

[14] Samarati, Pierangela, and Latanya Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, SRI International, 1998.

[15] Cox, Lawrence H. "Suppression methodology and statistical disclosure control." Journal of the American Statistical Association 75, no. 370 (1980): 377-385.

[16] Federal Committee on Statistical Methodology. Statistical policy working paper 22. Report on Statistical Disclosure Limitation Methodology, May 1994.

[17] Iyengar, Vijay S. "Transforming data to satisfy privacy constraints." In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 279-288. ACM, 2002.

[18] Meyerson, Adam, and Ryan Williams. "On the complexity of optimal k-anonymity." In Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pp. 223-228. ACM, 2004.

[19] Xu, Jian, Wei Wang, Jian Pei, Xiaoyuan Wang, Baile Shi, and Ada Wai-Chee Fu. "Utility-based anonymization using local recoding." In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 785-790. ACM, 2006.

[20] Gionis, Aristides, and Tamir Tassa. "k-Anonymization with minimal loss of information." Knowledge and Data Engineering, IEEE Transactions on 21, no. 2 (2009): 206-219.

[21] Byun, Ji-Won, Ashish Kamra, Elisa Bertino, and Ninghui Li. "Efficient k-anonymization using clustering techniques." In Advances in Databases: Concepts, Systems and Applications, pp. 188-200. Springer Berlin Heidelberg, 2007.

[22] Merz, Christopher J., and Patrick M. Murphy. "{UCI} repository of machine learning databases." (1998).

[23] Aggarwal, Gagan, Tomás Feder, Krishnaram Kenthapadi, Samir Khuller, Rina Panigrahy, Dilys Thomas, and An Zhu. "Achieving anonymity via clustering." In Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pp. 153-162. ACM, 2006.

[24] Jagannathan, Geetha, and Rebecca N. Wright. "Privacy-preserving distributed k-means clustering over arbitrarily partitioned data." In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, pp. 593-599. ACM, 2005.

[25] Pramanik, M. I., Lau, R. Y., & Yue, W. T. (2016). A Privacy Preserving Framework for Big Data in E-Government. PACIS 2016 Proceedings. Paper 72