A Clustering Approach for Achieving Data Privacy

Alina Campan, Traian Marius Truta, John Miller, and Raluca Sinca

Abstract — **New privacy regulations together with ever-increasing data availability and computational power have created a huge interest in data privacy research. One major research direction is built around *k*-anonymity property and its extensions, which are required for the released data. In this paper we present such an extension to *k*-anonymity, called *p*-sensitive *k*-anonymity, which solves some of the weaknesses that the *k*-anonymity model has been shown to have. We also introduce a new algorithm for enforcing *p*-sensitive *k*-anonymity on microdata sets based on a greedy clustering approach. To limit the amount of information loss the proposed algorithm uses cell-level generalization for categorical attributes and hierarchy-free generalization for numerical attributes. Our belief is that the above mentioned algorithm can be adjusted and used to enforce other similar privacy models as well, with better results than the algorithms originally proposed along with these models. Our experiments show that the proposed algorithm efficiently generates the masked microdata with *p*-sensitive *k*-anonymity property.**

I. Introduction

While the ever-increasing computational power together with the huge amount of individual data collected daily by various agencies is of great value for our society, they also pose a significant threat to individual privacy. Today, the privacy issue is not only on front pages of news agencies (the weeklong series "Privacy Lost", in [10]), but also the use and the disclosure of confidential information are subject to regulations in many countries. In the U.S., for example, privacy regulations promulgated by the Department of Health and Human Services as part of the *Health Insurance Portability and Accountability Act* (*HIPAA*) protect the confidentiality of electronic healthcare information [5]. Similar privacy regulations exist in other domains such as financial area [4]. More recently, Senator Hilary Rothman Clinton announced new legislation, the PROTECT (Privacy Rights and Oversight for Electronic and Commercial Transactions) Act of 2006 [12], that introduces new consumer privacy protections mechanisms.

A. Related Work

All these regulations together with the necessity of collecting personal information have funneled a huge interest in privacy research. Techniques to avoid the disclosure of confidential information exist in the literature [17]. Among them, the *k*-anonymity property required for the released *microdata* (datasets where each tuple belongs to an individual entity) was recently introduced [13] [14] and extensively studied [7] [15] etc. This property requires that in the released (a.k.a. masked) microdata every tuple will be indistinguishable from at least (*k*-1) other tuples with respect to a subset of attributes called key attributes or quasi-identifier attributes.

Recent results have shown that *k*-anonymity fails to protect the privacy of individuals in all situations [9] [15] [16] etc. New enhanced privacy models have been proposed in the literature to deal with *k*-anonymity limitations with respect to sensitive attribute disclosure. These models follow one of the following two approaches: the universal approach that uses the same privacy constraints for all individual entities, and the personalized approach that allows users or data owners to customize the amount of privacy needed for every individual. The first category of privacy protection models based on the universal approach includes: *p*-sensitive *k*-anonymity [15] with its improvement called extended *p*-sensitive *k*-anonymity [3], *l*-diversity [9], and (α, *k*)-anonymity [18]. The personalized privacy protection model we are aware of is personalized anonymity [16]. In this privacy model, a person can specify the degree of privacy protection for his/her sensitive values using a taxonomy provided by the data owner. The person will customize the level of desired privacy protection by choosing a node in this taxonomy [16].

B. Contributions

As mentioned, several privacy protection models based on the universal approach were introduced in the literature. Between them, we will focus on the *p*-sensitive *k*-anonymity model introduced in [15]. In addition to *k*-anonymity, this model requires each group of tuples with an identical combination of quasi-identifier attributes values, to maintain at least *p* distinct values for each confidential attribute (attribute which values must be protected) within the same group.

We introduce in this paper a method for anonymizing a microdata set such that its released version will satisfy *p*-sensitive *k*-anonymity. This method follows an approach found in [1] and [2], which consists in modeling and solving *k*-anonymization as a clustering problem. Namely, the algorithm takes an initial microdata set and establishes a partitioning of it into clusters. The released microdata set is

This work was supported by the Kentucky NSF EPSCoR Program under grant "*p*-Sensitive *k*-Anonymity Property for Microdata".

A. Campan (phone: 0040-746-881690; e-mail: alina@cs.ubbcluj.ro) and R. Sinca (e-mail: sr90938@linux.scs.ubbcluj.ro) are with the Department of Computer Science, Babes-Bolyai University, Cluj-Napoca, Romania.

T. M. Truta (e-mail: trutat1@nku.edu) and J. Miller (e-mail: millerj10@nku.edu) and are with the Department of Computer Science, Northern Kentucky University, USA.

formed by generalizing the quasi-identifier attributes of all tuples inside each cluster to the same values. The key element of this masking process is cluster formation. The clustering process is conducted such that the masked microdata produced in this manner will satisfy the *p*-sensitive *k*-anonymity requirement and the data utility lost by cluster-level tuple generalization will be minimized.

We evaluate the performance of our method by comparing the results it produces against the results provided by the non *p*-sensitive *k*-anonymization method presented in [2].

Currently, there are other algorithms proposed for enforcing models similar to *p*-sensitive *k*-anonymity on microdata [9] [18], however, they either use global recoding or consider only one sensitive attribute. Since our algorithm is based on local recoding (cluster-level generalization) and accepts multiple sensitive attributes, we believe it could lead to better results than these algorithms. Furthermore, with adequate adjustments, our algorithm could be used for enforcing (*α, k*)-anonymity or *l*-diversity on microdata as well.

The paper is structured as follows. Section 2 presents the *p*-sensitive *k*-anonymity model. Section 3 introduces the *GreedyPKClustering* algorithm for enforcing this privacy model on microdata. Experimental results are presented in Section 4. The paper ends with conclusions.

II. Privacy Protection for Sensitive Values

Let $\mathcal{IM}$ be the initial microdata and $\mathcal{MM}$ be the masked microdata. $\mathcal{IM}$ consists of a set of tuples over an attribute set. These attributes are classified into the following three categories:

- I_1, I_2,..., I_m are *identifier* attributes such as *Name* and *SSN* that can be used to identify a record.
- K_1, K_2,…, K_q are *key* or *quasi-identifier* attributes such as *ZipCode* and *Age* that may be known by an intruder.
- S_1, S_2,…, S_r are *confidential* or *sensitive* attributes such as *PrincipalDiagnosis* and *ICD9Code* that are assumed to be unknown to an intruder.

The identifier attributes are removed from the masked microdata, but the quasi-identifier and confidential attributes are usually released to the researchers. We assume that the values for the confidential attributes are not available from any external source. This assumption guarantees that an intruder can not use the confidential attribute values to increase his/her chances of disclosure. Unfortunately, an intruder may use record linkage techniques between quasi-identifier attributes and externally available information to glean the identity of individuals from the masked microdata. To avoid this possibility of disclosure, one frequently used solution is to modify the initial microdata; more specifically, the quasi-identifier attributes values, in order to enforce the *k*-anonymity property. After this change, the tuples from the resulting microdata can be clustered based on their common quasi-identifier attribute values.

Definition 1 (***QI-Cluster***): A ***QI-cluster*** consists of the tuples with identical combination of **quasi-identifier** attribute values in a given microdata.

There is no consensus in the literature over the term used to denote a *QI-cluster*. This term was not defined when *k*-anonymity was introduced [13] [14]. More recent papers use different terminology such as *equivalence class* [18] and *QI-group* [16]. Since our algorithm is based on clustering concepts, we decided to use *QI-cluster* term in this paper (or simply cluster, when there is no danger of confusion). Now we can rigorously define *k*-anonymity based on the minimum size of all *QI-clusters*.

Definition 2 (***K-Anonymity Property***): The ***k-anonymity property*** for a $\mathcal{MM}$ is satisfied if every *QI-cluster* from a $\mathcal{MM}$ contains *k* or more tuples.

Unfortunately, as pointed out in the literature [9] [15] [18], *k*-anonymity does not provide the amount of confidentiality required for every individual. To briefly justify this affirmation, we distinguish between two possible types of disclosure; namely**,** identity disclosure and attribute disclosure. *Identity disclosure* refers to re-identification of an entity (person, institution) and *attribute disclosure* occurs when the intruder finds out something new about the target entity [6]. *K*-anonymity protects against identity disclosure but fails to protect against attribute disclosure when all tuples of a *QI-cluster* share the same value for one sensitive attribute [15]. This attack is called *homogeneity attack* [9] and can be avoided by enforcing a more powerful anonymity model than *k*-anonymity, for example *p*-sensitive *k*-anonymity. A different type of attack, called *background attack*, is presented in [9]. In this scenario, the attacker has some background information that allows him / her to rule out some possible values of the sensitive attributes for specific individuals. Protection against background attacks is more difficult since the type of background knowledge is unknown by the data owner. To be certain of success in case of a background attack, particular assumptions should be made, and anonymization techniques by themselves will not fully solve this problem [18]. Still, existing techniques try to perform as well as possible in case of background attacks.

We present next the *p*-sensitive *k*-anonymity model (introduced in [15]), one of several similar privacy models, recently proposed, that guard against attribute disclosure.

A. P-Sensitive K-Anonymity And Its Extension

The *p*-sensitive *k*-anonymity model is a natural extension of *k*-anonymity. Both the simple [15] and extended [3] versions consider several sensitive attributes that must be protected against attribute disclosure. Although initially designed to protect against homogeneity attacks, it also performs well against different types of background attacks. It has the advantage of simplicity and allows the data owner to customize the desired protection level by setting various values for p and k. Intuitively, the larger the parameter p, the better is the protection against both types of attribute disclosure attacks.

Definition 3 (***p-Sensitive k-Anonymity Property***): A $\mathcal{MM}$ satisfies ***p-sensitive k-anonymity property*** if it satisfies k-anonymity and the number of distinct attributes for each confidential attribute is at least p within the same *QI-cluster* from the $\mathcal{MM}$.

Fig. 1 illustrates how a masking process can protect data against identity and attribute disclosure. The first dataset (Fig. 1.a) represents the initial microdata, from which identifier attributes values have been removed. An intruder can link this dataset with the external data in Fig. 1.b), to identify individuals and to find new information about them. However, if initial microdata is generalized as in Fig. 1.c), all individuals are protected against identity and attribute disclosure. Masked microdata in Fig. 1.c) is 2-sensitive (w.r.t. sensitive attribute *Illness*) and 3-anonymous (w.r.t. quasi-identifiers *Age*, *Zip*, and *Gender*).

Age	Zip	Gender	Illness
25	41076	Male	Diabetes
25	41075	Male	Heart Disease
27	41076	Male	Diabetes
35	41099	Female	Colon Cancer
38	48201	Female	Breast Cancer
36	41075	Female	HIV

a) initial microdata

Name	Age	Zip	Gender
Sam	25	41076	Male
Eric	25	41075	Male
Sandra	35	41099	Female
Gloria	38	48201	Female
Tanisha	36	41075	Female
Don	27	41076	Male

b) external data

Age	Zip	Gender	Illness
20-30	4107*	Male	Diabetes
20-30	4107*	Male	Heart Disease
20-30	4107*	Male	Diabetes
30-40	4****	Female	Colon Cancer
30-40	4****	Female	Breast Cancer
30-40	4****	Female	HIV

c) masked microdata

Fig. 1. A masking process

The p-sensitive k-anonymity property can not always be enforced to a microdata set $\mathcal{IM}$. Next, we give necessary conditions that express when this is possible [15].

Let s_j be the number of distinct values for each confidential attribute S_j (j = 1..r), in $\mathcal{IM}$. In this case p must always be less than or equal to $\min_{j=1,r}(s_j)$. For instance, if we consider *Sex* as a confidential attribute, because the number of distinct values for *Sex* is only two, the maximum possible value for p is two. Therefore, the following condition must hold in order to form a p-sensitive k-anonymous masked microdata $\mathcal{MM}$ from a given initial microdata $\mathcal{IM}$.

Condition 1 (*First necessary condition for an $\mathcal{MM}$ to have p-sensitive k-anonymity property*): The minimum number of distinct values for each confidential attribute in $\mathcal{IM}$ must be greater than or equal to p.

A second necessary condition establishes the maximum possible number of *QI-clusters* in the masked microdata that satisfy p-sensitive k-anonymity. To specify this upper bound we use the maximum between cumulative descending ordered frequencies for each sensitive attribute (labeled as cf_i) in $\mathcal{IM}$ [15].

Condition 2 (*Second necessary condition for a $\mathcal{MM}$ to have p-sensitive k-anonymity property*): The maximum possible number of *QI-clusters* in the masked microdata is $pMaxGroups = \min_{i=1,p} \left\lfloor \frac{n - cf_{p-i}}{i} \right\rfloor$.

Sometimes the domain of the confidential attributes, especially the categorical ones, can be organized according to some hierarchies. For example, in medical datasets the *Illness* attribute has values as specified by the *ICD9* codes [3]. The data owner may want to protect not only the leaf values as in the p-sensitive k-anonymity model, but also values found at higher levels. For example, the information that a person has cancer (not a leaf value in this case) needs to be protected regardless of the cancer type (colon cancer, prostate cancer, breast cancer are examples of leaf nodes in this hierarchy). If p-sensitive k-anonymity property is enforced for the released microdata, it is possible that all of the *Illness* attribute values could be descendants of the cancer node in the corresponding hierarchy for one *QI-cluster*; therefore, leading to disclosure. To avoid such situations, the *extended p-sensitive k-anonymity* model was introduced in [3]. As shown there, any method that can be used to enforce p-sensitive k-anonymity to a microdata set can also be used to enforce the extended model version to that microdata set. Henceforth, the *GreedyPKClustering* algorithm presented next can be used to enforce both simple and extended p-sensitive k-anonymity model versions.

III. A Greedy Clustering Algorithm for Achieving *P*-Sensitive *K*-Anonymity

The algorithm described in this section, called the *GreedyPKClustering* algorithm, performs a greedy clustering processing to impose p-sensitive k-anonymity to a microdata set $\mathcal{IM}$.

First, the algorithm establishes a "good" partitioning of all tuples from $\mathcal{IM}$ into clusters. Next, all tuples within each cluster are made uniform w.r.t. the quasi-identifier attributes; this homogenization is achieved by using (as many of the existing k-anonymization algorithms do) quasi-identifier attributes generalization.

In order for the two requirements of the p-sensitive k-anonymity model to be fulfilled, each cluster has to contain at least k tuples and at least p different values for every confidential attribute. Consequently, a first criterion to lead the clustering process is to ensure each cluster has enough diversity w.r.t. the confidential attributes (the p-sensitive requirement), followed by enough (at least k) elements. As it is well known, attribute generalization results in information loss; therefore, a second criterion used during clustering is to minimize an information loss cost metrics between initial and released microdata, caused by the subsequent cluster-level quasi-identifier attributes generalization.

To sum up, in order to obtain good quality masked microdata, the clustering algorithm uses two measures: one for cluster diversity and one for information loss, which correspond to the two criteria explained above. We introduce next the cluster diversity and information loss measures we used. Then, the clustering algorithm will be presented and explained.

A. *Cluster Diversity And Information Loss*

Let X^i, i=1..n, be the tuples from $\mathcal{IM}$ subject to p-sensitive k-anonymization. We denote an individual tuple as $X^i = \{k_1^i, k_2^i, ..., k_q^i, s_1^i, s_2^i, ..., s_r^i\}$, where k^i s are the values for the quasi-identifier attributes and s^i s are the values for the confidential attributes.

Definition 4: The ***diversity of two tuples***, X^i and X^j w.r.t. the confidential attributes is given by:

$$diversity(X^i, X^j) = \sum_{l=1}^{r} w_l \cdot \delta(s_l^i, s_l^j)\text{, where}$$

$$\delta(s_l^i, s_l^j) = \begin{cases} 1, \text{ } if \text{ } s_l^i <> s_l^j \\ 0, \text{ } if \text{ } s_l^i = s_l^j \end{cases} \text{ and } \sum_{l=1}^{r} w_l = 1 \text{ are the weights of}$$

the sensitive attributes.

The data owner can choose different criteria to define this weights vector. One good selection of the weight values is to initialize them as inversely proportional to the number of distinct sensitive attribute values in the microdata $\mathcal{IM}$. Along the entire paper we use this choice for the weights in all the experiments.

Definition 5: The ***diversity between a tuple*** X^i ***and a cluster*** cl is given by $diversity(X^i, cl) = \sum_{l=1}^{r} w_l \cdot \rho(s_l^i, cl_l)$,

where:

$$\rho(s_l^i, cl_l) = \begin{cases} 1, \text{ if } s_l^i \text{ does not exist between the } S_l \text{ values in } cl \\ 0, \text{ if } s_l^i \text{ exists between the } S_l \text{ values in } cl \end{cases}$$

and $\sum_{l=1}^{r} w_l = 1$ have the same meaning as in **Definition 4**.

For the information loss we use the measures introduced in [2]. The loss of information occurs due to the generalization of quasi-identifier attributes. For categorical attributes we used generalization based on predefined hierarchies at the cell level [8]. We denote by H_C the hierarchies (domain and value) associated to the categorical quasi-identifier attribute C. For numerical attributes we use the hierarchy-free generalization [7], which consists of replacing the set of values to be generalized with the smallest interval that includes all the initial values. We generalize each cluster to the least general tuple that represents all tuples in that group [3]. We call generalization information for a cluster the minimal covering tuple for that cluster, and we define it as follows. (Of course, generalization and coverage refer only to the quasi-identifier part of the tuples).

Definition 6: Let $cl = \{r_1, r_2, \ldots, r_u\}$ be a cluster of tuples selected from $\mathcal{IM}$, KN = $\{N_1, N_2, ..., N_s\}$ be the set of numerical quasi-identifier attributes and KC = $\{C_1, C_2, \ldots, C_t\}$ be the set of categorical quasi-identifier attributes. The ***generalization information of cl***, w.r.t. quasi-identifier attribute set K = KN $\cup$ KC is the "tuple" $gen(cl)$, having the scheme K, where:

- For each categorical attribute $C_j \in$ K, $gen(cl)[C_j]$ = the lowest common ancestor in H_{Cj} of $\{r_1[C_j], \ldots, r_u[C_j]\}$;
- For each numerical attribute $C_j \in$ K, $gen(cl)[C_j]$ = the interval $[\min\{r_1[C_j], \ldots, r_u[C_j]\}, \max\{r_1[C_j], \ldots, r_u[C_j]\}]$.

For cluster cl, its generalization information $gen(cl)$ is the tuple having as value for each quasi-identifier attribute, numerical or categorical, the most specific common generalized value for all that attribute values from cl tuples. In a $\mathcal{MM}$, each tuple from cluster cl will have its quasi-identifier attributes values replaced by $gen(cl)$.

Now we have all the tools to introduce information loss measures.

Definition 7: Let cl be a cluster, $gen(cl)$ its generalization information, and K = $\{N_1, N_2, .., N_s, C_1, C_2, .., C_t\}$ the set of quasi-identifier attributes. The ***information loss*** caused by generalizing quasi-identifier attributes of the cl tuples to $gen(cl)$ is:

$$IL(cl) = |cl| \cdot \left(\sum_{j=1}^{s} \frac{size(gen(cl)[N_j])}{size\left(\left[\min_{r \in IM} r[N_j], \max_{r \in IM} r[N_j]\right]\right)} + \sum_{j=1}^{t} \frac{height(\Lambda(gen(cl)[C_j]))}{height(H_{C_j})} \right)$$

where:

- $|cl|$ denotes the cluster cl cardinality;
- $size([i_1, i_2])$ is the size of the interval $[i_1, i_2]$ $(i_2 - i_1)$;
- $\Lambda(w)$, $w \in H_{Cj}$ is the subhierarchy of H_{Cj} rooted in w;
- height(H_{Cj}) denotes the height of the tree hierarchy H_{Cj}.

Definition 8: ***Total information loss*** for a partition $\mathcal{S} = \{cl_1, cl_2, \ldots, cl_v\}$ of the microdata set is the sum of the information loss measure for all the clusters in $\mathcal{S}$.

B. *GreedyPKClustering Algorithm*

Using the above introduced measures, in this section, we explain how clustering is performed for a given initial microdata set $\mathcal{IM}$.

The *QI-clusters* are formed one at a time. For forming one *QI-cluster*, a tuple in $\mathcal{IM}$ not yet allocated to any cluster is selected as a seed for the new cluster. Then the algorithm gathers tuples to this currently processed cluster until it satisfies both requirements of the p-sensitive k-anonymity model. At each step, the current cluster grows with one tuple.

This tuple is selected, of course, from the tuples not yet allocated to any cluster. If *p*-sensitivity part is not yet satisfied, then the chosen tuple is the one most probable to enrich the diversity of the current cluster with regard to the confidential attributes values. This selection is made by the diversity measure between a tuple and a cluster. If the *p*-sensitivity part is already satisfied for every confidential attribute, then the least different or diverse tuple (w.r.t. the confidential attributes) of the current cluster is chosen. This selection is justified by the need to spare other different confidential values, not present in the current cluster, in order to be able to form as many as possible new *p*-sensitive clusters. When a tie happens, i.e. multiple candidate tuples exist conforming to the previous selection criteria, then the tuple that minimizes the cluster's *IL* growth will be preferred.

It is possible that the last constructed cluster will contain less than *k* tuples or it will not satisfy *p*-sensitivity requirement. In that case, this cluster needs to be dispersed between the previously constructed groups. Each of its tuples will be added to the cluster whose *IL* will minimally increase by that tuple addition.

The ***GreedyPKClustering*** algorithm is shown in Fig. 2.

```
Algorithm GreedyPKClustering is
Input IM – microdata;
        $p$, $k$ – as in $p$-sensitive $k$-anonymity;
Output S={$cl_1$,$cl_2$,…,$cl_v$}; $\bigcup_{j=1}^{v} cl_j$ = IM; $cl_i \bigcap cl_j$ =∅,
        $i$,$j$=1..$v$, $i$≠$j$; |$cl_j$|≥$k$, $cl_j$ is $p$-sensitive
        for every $S^1$, $j$=1..$v$ – a set of clusters
        that ensures $p$-sensitive $k$-anonymity;
S = ∅; $i$ = 1;
$r_{seed}$ = a randomly selected tuple from IM;
Repeat
   $cl_i$ = ∅;
   $r_{seed} = \underset{r\in IM}{\text{argmax}}\ diversity(r_{seed}, r)$ ;
   // the most diverse tuple from IM wrt. old $r_{seed}$
   $cl_i$ = $cl_i$ ∪ {$r_{seed}$};
   IM = IM – {$r_{seed}$};
   Repeat
      $r^* = \underset{IL}{\text{argmin}}\ \underset{r\in IM}{\text{argmax}}\ diversity(r, cl_i)$ ;
      // the tuple within the most diverse tuples
      // w.r.t. $cl_i$ that produces the minimal $IL$
      // growth when added to $cl_i$
      $cl_i$ = $cl_i$ ∪ {$r^*$};
      IM  = IM – {$r^*$};
   Until ($cl_i$ is $p$-sensitive) or (IM = ∅);
   If (|$cl_i$| < $k$) and (IM ≠ ∅) then
      Repeat
         $r^* = \underset{IL}{\text{argmin}}\ \underset{r\in IM}{\text{argmin}}\ diversity(r, cl_i)$ ;
         // the tuple within the least diverse
         // tuples w.r.t. $cl_i$ that produces the
         // minimal $IL$ growth when added to $cl_i$
         $cl_i$ = $cl_i$ ∪ {$r^*$};
         IM = IM – {$r^*$};
      Until ($cl_i$ is $k$-anonymous) or (IM = ∅);
   End If;
   If (|$cl_i$| ≥ $k$ and $cl_i$ is $p$-sensitive) then
      S = S ∪ {$cl_i$}; $i$++;
   Else
      DisperseCluster(S, $cl_i$);
      // this happens only for the last cluster
   End If;
Until IM = ∅;
End GreedyPKClustering.

Function DisperseCluster(S, $cl$)
   S = S – {$cl$};
   For every $r$ ∈ $cl$ do
      $cl_{u^*}$ = FindBestCluster($r$, S);
      $cl_{u^*}$ = $cl_{u^*}$ ∪ {$r$};
   End For;
End DisperseCluster;

Function FindBestCluster($r$, S) is
   bestCluster = null;
   infoLoss = ∞;
   For every $cl_j$ ∈ S do
      If $IL$($cl_j$ ∪ {$r$}) < infoLoss   then
         infoLoss = $IL$($cl_j$ ∪ {$r$});
         bestCluster = $cl_j$;
      End If;
   End For;
   Return bestCluster;
End FindBestCluster;
```

Fig. 2. The GreedyPKClustering Algorithm

IV. Experimental Results

In our experiments we used data from the Adult database from the UC Irvine Machine Learning Repository [11]. This database has become the benchmark for *k*-anonymity algorithms [7]. The experiments reported in [2] are also based on it, and in this section we compare, in terms of efficiency and results quality, the non *p*-sensitive *k*-anonymization algorithm from [2] (called *greedy_k_member_clustering*) with our algorithm. We chose this algorithm for comparison, because it is based on the same clustering approach as *GreedyPKClustering* and both use local recoding – i.e. generalization at cluster-level. We intend to extend our experiments and perform comparative tests with other algorithms proposed to enforce models equivalent with *p*-sensitive *k*-anonymity (*l*-diversity and (*α*, *k*)-anonymity). However, we think that an algorithm based on global recoding will produce weaker results (in terms of *IL*) compared to our local recoding algorithm, and this without connection to the enforced anonymity model.

The algorithms we tested have been implemented in Java, and tests were executed on a single processor machine running Windows XP with 2.26 GHz and 256 MB of RAM.

A set of experiments has been conducted for an *IM* consisting in 10000 tuples randomly selected from the adult dataset. In all the experiments, we considered *age, workclass, marital-status, race*, *sex* and *native-country* as the set of quasi-identifier attributes; and *education_num*, *education* and *occupation* as the set of confidential attributes. Microdata *p*-sensitive *k*-anonymity was enforced in respect to the quasi-identifier consisting of all 6 quasi-identifier attributes and all 3 confidential attributes.

For the described *IM*, we applied the *GreedyPKClustering* algorithm with different values of *k* and *p*. For every value of *k* we run the *k*-anonymization algorithm (without *p*-sensitive guarantee) presented in [2].

Fig. 3 presents comparatively the total *IL* of the *QI-clusters* the two algorithms, *GreedyPKClustering* and *greedy_k_member_clustering*, produced for different *k* values. The *IL* measure for a cluster increases with the cluster cardinality and the diversity of quasi-identifier attributes values of the cluster elements. This fact explains why total *IL* for *MM* generally increases with *k*, for both *greedy_k_member_clustering* and *GreedyPKClustering* algorithms (Fig. 3). Also, the *IL* increases with *p*. The larger *p* value is, the smaller is the number of *p*-sensitive *QI-clusters* that can be formed for the same microdata. Of course, the *QI-clusters* are, consequently, larger and more diverse w.r.t. the quasi-identifier attributes as well, so they are also characterized by larger *IL* values. The difference of *IL* results between the two algorithms, for the same *k* values, is justified by this argument (Fig. 3). However, some of the *QI-clusters* produced by *greedy_k_member_ clustering* violate the *p*-sensitivity condition and expose microdata to attribute disclosure. For a fixed *k* value, as *p* increases, the number of *QI-clusters* that violate *p*-sensitive requirement increases, as can be seen in Table 1.

The *IL* decrease between *p*=2 and *p*=4 for all *k* values, when using *GreedyPKClustering*, can be explained as follows. During a *QI-cluster* formation, after *p*-sensitive property has been reached for it, and until it numbers *k* tuples, the cluster growth is performed by adding new tuples that minimize cluster diversity increase (as already explained in Section 3.2). This seems to correspond, for some datasets, to adding elements that increase cluster *IL*. The smaller *p* is compared to *k*, earlier is *p*-sensitivity reached for a cluster, and more intensively the mentioned cluster growth condition is used. When *p* is closer to *k*, *p*-sensitivity is not reached often for a cluster before it has *k* elements, so this cluster growth condition is rarely used. As a consequence, for some data sets it could be more appropriate to select the tuples that minimize cluster *IL* to be included in the *QI-cluster*, after *p*-sensitive requirement is reached. Still more research is needed to understand how this modified criteria would have an effect on the total number of *QI-clusters*.

Fig. 4 presents comparatively the execution time of *GreedyPKClustering* and *greedy_k_member_clustering* algorithms for different *k* values. As depicted in these figures, *GreedyPKClustering* runs faster than *greedy_k_ member_clustering* for the same *k* value and any *p* value.

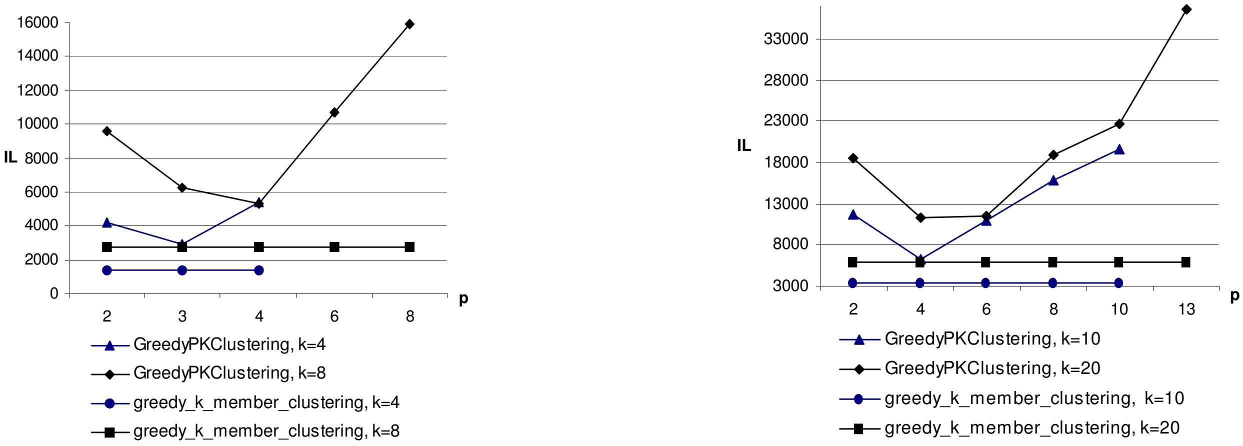

Fig. 3. IL for *GreedyPKClustering* and *greedy_k_member_clustering*, for different *k* and *p* values

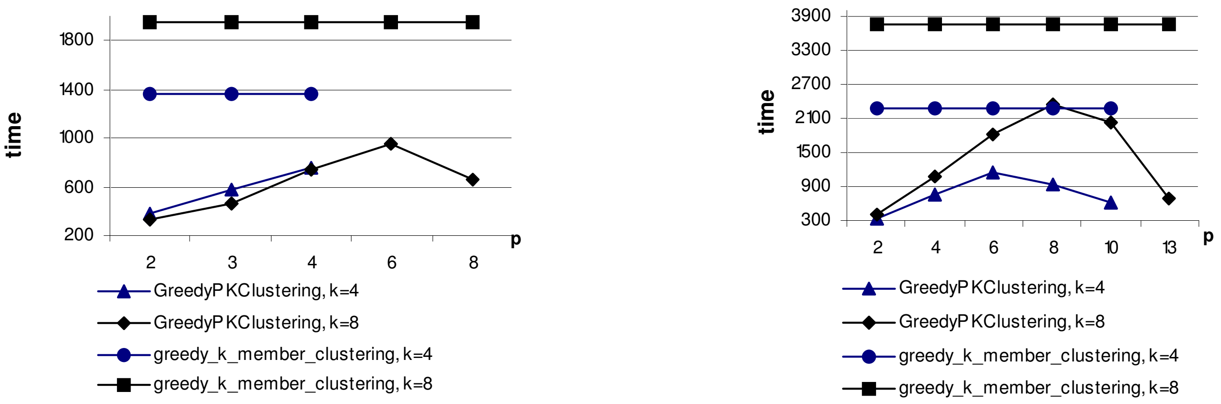

Fig. 4. Execution Time for *GreedyPKClustering* and *greedy_k_member_clustering*, for different *k* and *p* values

This is caused by the fact that cluster *IL* calculation requires more computational effort than cluster diversity calculation.

In the same time, *QI-clusters* formation is guided in *GreedyPKClustering* mainly in respect to cluster diversity improvement, while the guiding criterion used in *greedy_k_member_clustering* is cluster *IL*.

TABLE 1.
VIOLATIONS OF THE *P*-SENSITIVE REQUIREMENT BY *GREEDY_K_MEMBER_CLUSTERING* RESULTS

k		k=4	k=8	k=10	k=20
No of clusters produced for current *k*		2500	1250	1000	500
No of clusters produced for current *k* that violate *p*-sensitivity	p=2	94	1	0	0
	p=3	1048	-	-	-
	p=4	2308	384	147	18
	p=6	-	1205	852	156
	p=8	-	1250	999	431
	p=10	-	-	1000	497
	p=20	-	-	-	500

The *GreedyPKClustering* algorithm produces good results w.r.t. the number of *p*-sensitive *QI-clusters* it generates. As expressed by **Condition 2**, the confidential attribute values distributions strongly influence the number of *p*-sensitive *QI-clusters* that can be formed. We detail in Table 2 the maximum numbers of *p*-sensitive *QI-clusters* that can be formed in the microdata set $\mathcal{IM}$ we considered.

TABLE 2.
PMAXGROUPS FOR THE 3 SENSITIVE ATTRIBUTES

P	2	3	4	6	8	10	13
pMaxGroups	6721	3361	2225	926	556	365	47

Table 3 reports the cardinality of partitions created by *GreedyPKClustering* for different *k* and *p* values, and compared to the maximal possible number of *p*-sensitive *QI-clusters*, given in Table 2. Notice: **Condition 2** provides a superior limit of the number of *p*-sensitive *QI-clusters* that can be formed in a microdata set, but not the actual number of such clusters that exist in data. So, even the optimal partition w.r.t. the partition cardinality criterion could consist in less *p*-sensitive *QI-clusters* than the number estimated by **Condition 2**.

TABLE 3.
POSSIBLE NO OF CLUSTERS FOR CURRENT *K*, *P* VS ACTUAL NO OF CLUSTERS FOR CURRENT *K*, *P*

k / poss. no. clusters for *k*	4 / 2500	8 / 1250	10 / 1000	20 / 500
p=2	2500 / 2500	1250 / 1250	1000 / 1000	500 / 500
p=3	2500 / 2473	-	-	-
p=4	2225 / 1987	1250 / 1241	1000 / 999	500 / 500
p=6	-	926 / 811	926 / 759	500 / 493
p=8	-	556 / 456	556 / 444	500 / 372
p=10	-	-	365 / 259	365 / 244
p=13	-	-	-	47 / 46

V. CONCLUSIONS AND FUTURE WORK

In this paper, a clustering algorithm to generate masked microdata with *p*-sensitive *k*-anonymity property was introduced. Our experiments have shown that the proposed algorithm generates the masked microdata in an efficient manner. Although, as expected, the information loss is increased due to the "*p*-sensitive" condition (the increase is gradual) and for small values of *p* the information loss is comparable with the one obtained using algorithms [2] that generate *k*-anonymized masked microdata sets.

We believe with adequate diversity/information loss measures, this algorithm could be used for enforcing (*α, k*)-anonymity or *l*-diversity on microdata as well. As our algorithm is based on local recoding (cluster-level generalization) and accepts multiple sensitive attributes, it could lead to better results than the currently proposed algorithms in [9] [18], which either use global recoding or consider only one sensitive attribute.

REFERENCES

[1] Aggarwal, G., Feder, T., Kenthapadi, K., Motwani, R., Panigrahy, R., Thomas, D., Zhu, A., "Achieving Anonymity via Clustering", Proc. of the 25th ACM PODS, pp. 153-162, 2006.

[2] Byun, J.W., Kamra, A., Bertino, E., Li, N., "Efficient *k*-Anonymity using Clustering Techniques", CERIAS Technical Report 2006-10, 2006.

[3] Campan, A., Truta, T.M., "Extended *p*-Sensitive *k*-Anonymity for Privacy Protection", in Studia Universitatis Babes-Bolyai, Informatica, Vol. LI(2), pp. 19-30, 2006.

[4] GLB, "Gramm-Leach-Bliley Financial Services Moderniza-tion Act", available online at http://banking.senate.gov/conf/, 1999.

[5] HIPAA, "Health Insurance Portability and Accountability Act", available online at http://www.hhs.gov/ocr/hipaa, 2002.

[6] Lambert, D., "Measures of Disclosure Risk and Harm", Journal of Official Statistics, Vol. 9, 313-331, 1993.

[7] LeFevre, K., DeWitt, D., Ramakrishnan, R., "Mondrian Multidimensional *k*-Anonymity", Proc. of the IEEE International Conference of Data Engineering, Atlanta, 2006.

[8] Lunacek, M., Whitley, D., Ray, I., "A Crossover Operator for the *k*-Anonymity Problem", Proc. of GECCO 2006, 1713 – 1720.

[9] Machanavajjhala A., Gehrke J., Kifer D., "*l*-diversity: privacy beyond *k*-anonymity", Proc. of the 22nd IEEE International Conference on Data Engineering, Atlanta, 2006.

[10] MSNBC, "Privacy Lost", available online at http://www.msnbc.msn.com/id/15157222/, 2006.

[11] Newman, D.J., Hettich, S., Blake, C.L., Merz, C.J., "UCI Repository of Machine Learning Databases", available online at www.ics.uci.edu/~mlearn/MLRepository.html, University of California, Irvine, 1998.

[12] PROTECT, "Privacy Rights and Oversight for Electronic and Commercial Transactions Act", available online at http://www.theorator.com/bills109/s3713.html, 2006.

[13] Samarati, P., "Protecting Respondents Identities in Microdata Release", IEEE Transactions on Knowledge and Data Engineering, Vol. 13, No. 6, 1010-1027, 2001.

[14] Sweeney, L., "*k*-Anonymity: A Model for Protecting Privacy", International Journal on Uncertainty, Fuzziness, and Knowledge-based Systems, Vol. 10, No. 5, pp. 557 – 570, 2002.

[15] Truta, T.M., Bindu, V., "Privacy Protection: *p*-Sensitive *k*-Anonymity Property", Workshop on Privacy Data Management, 22th IEEE Intl. Conference of Data Engineering, Atlanta, 2006.

[16] Xiao, X., Tao, Y., "Personalized Privacy Preservation", Proc. of the ACM SIGMOD, Chicago, Illinois, pp. 229-240, 2006.

[17] Willemborg, L., Waal, T. (ed), *Elements of Statistical Disclosure Control*, Springer Verlag, 2001.

[18] Wong, R.C-W., Li, J., Fu, A. W-C., Wang, K., "(*α, k*)-Anonymity: An Enhanced *k*-Anonymity Model for Privacy-Preserving Data Publishing", Proc. of the ACM KDD, 2006.