A COMPARISON BETWEEN LOCAL AND GLOBAL RECODING ALGORITHMS FOR ACHIEVING MICRODATA *P*-SENSITIVE *K*-ANONYMITY

Traian Marius Truta, Alina Campan,
Michael Abrinica, John Miller

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, which is required for the released data. Although k-anonymity protects against identity disclosure, it fails to provide an adequate level of protection with respect to attribute disclosure. We introduced a new privacy protection property called p-sensitive k-anonymity that avoids this shortcoming. We developed new algorithms (*GreedyPKClustering* and *EnhancedPKClustering*) and adapted an existing algorithm (*Incognito*) to generate masked microdata with p-sensitive k-anonymity property. All these algorithms try to reduce the amount of information lost while transforming data to conform to p-sensitive k-anonymity. They are different in the masking methods they use. The new algorithms are based on local recoding masking methods. *Incognito*, initially designed for k-anonymity, uses global recoding for masking. This paper's goal is to compare the impact of the masking method on the quality of the masked microdata obtained. For this we compare the quality of the results (cost measures based on data utility) and the efficiency (running time) of these three algorithms for masking both real and synthetic data sets.

2000 *Mathematics Subject Classification*: 68P15, 68U35.

1. Introduction

The ever-increasing storage and computational power resulted in the accumulation of large datasets at different companies. Many of these datasets contain private information about individuals or other entities, information that must be protected from, both, a moral point of view and according to several new privacy regulations ([7], [6], etc.). At the same time, appropriate analysis of the respective datasets can generate useful knowledge. However, data analysis is usually performed not only by the data owners, but also by third-parties, who should not have access to those private data. In order to cope with the data privacy requirements and the need for data analysis, the data must be, before its release to third-parties, subject to a transformation (or so-called masking) process. This process will ensure privacy while preserving the data's information and knowledge content as much as possible.

It is worth noting that there are an increasing number of privacy regulations in many countries. For instance, the *Privacy Rule* section from the *Health Insurance Portability and Accountability Act* (*HIPAA*) is one of the well-known privacy regulations in the U.S. that protects the confidentiality of electronic healthcare information [7]. Similar privacy regulations exist in other domains ([6], [2]). Recently, Senator Hilary Rothman Clinton called for a comprehensive privacy agenda: a Privacy Bill of Rights that introduces new consumer privacy protection mechanisms. Senator Clinton announced that she will develop legislation to enact this Bill of Rights, called the Privacy Rights and Oversight for Electronic and Commercial Transactions Act of 2006, the PROTECT Act [16]. In Romania, the right of privacy is recognized in the Constitution, Article 26 [15]. Recently, the Parliament enacted Law No. 676/2001 on the Processing of Personal Data and the Protection of Privacy in the Telecommunications Sector and Law No. 677/2001 for the Protection of Persons concerning the Processing of Personal Data and the Free Circulation of Such Data [15]. These laws show that there is an increasing concern in Romania to align with European Union privacy regulations.

One major research direction in the field of data privacy is built around the k-anonymity property, which is required for the released data. A microdata set (a dataset where each tuple corresponds to one individual) conforms to this property if every tuple within it is indistinguishable from at least (k-1) other tuples, with respect to a set of attributes called quasi-identifier or key attributes. Many research efforts have been directed towards finding methods to anonymize datasets using k-anonymity property ([1], [3], [9], etc.).

However, recent results have shown that k-anonymity fails to protect the privacy of individuals in all situations ([20], [13], [25], etc.). Several privacy models have been proposed in the literature to avoid k-anonymity shortcomings: p-sensitive k-anonymity [20] with its extension called extended p-sensitive k-anonymity [5], l-diversity [13], (α, k)-anonymity [24], and t-closeness [11]. They were accompanied by algorithmic solutions for transforming (or masking) the datasets in order to conform to those models.

Contributions: Most of the anonymization algorithms, those dedicated to k-anonymity masking, as well as those that mask the data according to the extended models, use mainly two general data transformation techniques to attain their aim. These techniques are data generalization (also known as data recoding, where each masked value is faithful to the original) and data suppression, the latter being more rarely used. This paper's goal is to compare the impact of the masking method used on the quality of the masked microdata obtained. With that end in view, we present an extensive comparison between three algorithms used to create p-sensitive k-anonymous datasets: *Incognito* [9], *GreedyPKClustering* [5], and *EnhancedPKClustering* [22]. Although all these algorithms try to reduce information loss while transforming data to conform to p-sensitive k-anonymity, they are different in the masking methods they use. The *GreedyPKClustering* and *EnhancedPKClustering* are based on local recoding masking methods. *Incognito*, initially designed for k-anonymity, uses global recoding for masking. We compare the results' quality (cost measures based on data utility) and the efficiency (running time) of these three algorithms for masking both real and synthetic data sets, and draw a conclusion.

The structure of the paper will be as follows. The k-anonymity and p-sensitive k-anonymity privacy models along with their properties will be presented in Section 2. Section 3 describes different generalization techniques widely used for data anonymization. Section 4 contains a brief description of the above mentioned algorithms: *Incognito* [9], *GreedyPKClustering* [5], and *EnhancedPKClustering* [22]. The cost measures used for comparing masked datasets' quality are also reported in Section 4. Section 5 reports experimental results for the three algorithms, applied on real and synthetic datasets. The paper ends with conclusions and a bibliography.

2. Data Anonymity Models

Let $\mathcal{IM}$ be the initial microdata and $\mathcal{MM}$ be the released (a.k.a. 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. These attributes are present only in the initial microdata because they express information which can lead to a specific entity.
- K_1, K_2, …, K_n are key or quasi-identifier attributes such as *ZipCode* and *Age* that may be known by an intruder. Quasi-identifier attributes are present in the masked microdata as well as in the initial microdata.
- S_1, S_2, …, S_r are confidential or sensitive attributes such as *PrincipalDiagnosis* and *Income* that are assumed to be unknown to an intruder. Confidential attributes are present in the masked microdata as well as in the initial microdata.

While the identifier attributes are removed from the published microdata, the quasi-identifier and confidential attributes are usually released to the researchers/analysts. A general assumption, as noted, is 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 attributes values to increase his/her chances of disclosure, and, therefore, modifying these attributes' values is unnecessary. Unfortunately, an intruder may use record linkage techniques [23] between quasi-identifier attributes and external 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 or one of its extensions.

For example, let $\mathcal{IM}$ be the dataset in Table 1, where *Name* and *SSN* are the identifier attributes, *Age* and *Zip* are the quasi-identifier attributes, and *Diagnosis* and *Income* represent the sensitive attributes. Having the identifier attributes removed, the resulted dataset is obviously not 2-sensitive, as its third tuple can be clearly associated to the person named Charley by an intruder who knows that his target is a 44-aged person living in 48201 area code. By generalizing the zip code af the first 3 tuples in $\mathcal{IM}$ as depicted in Table 2, the obtained dataset $\mathcal{MM}$ is 2-anonymous.

In order to rigorously and succinctly express k-anonymity and p-sensitive k-anonymity properties, we use the following concept:

Definition 1 ***(QI-cluster)****:* Given a microdata, a QI-cluster consists of all the tuples with identical combination of quasi-identifier attribute values in that microdata.

Table 1: An $\mathcal{IM}$ dataset

Name	SSN	Age	Zip	Diagnosis	Income
Alice	123456789	44	48202	AIDS	17,000
Bob	323232323	44	48202	AIDS	68,000
Charley	232345656	44	48201	Asthma	80,000
Dave	333333333	55	48310	Asthma	55,000
Eva	666666666	55	48310	Diabetes	23,000

Table 2: A 2-anonymous $\mathcal{MM}$ dataset corresponding to $\mathcal{IM}$

Age	Zip	Diagnosis	Income
44	4820*	AIDS	17,000
44	4820*	AIDS	68,000
44	4820*	Asthma	80,000
55	48310	Asthma	55,000
55	48310	Diabetes	23,000

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 ([17], [18]). More recent papers use different terminologies such as *equivalence class* [24] and *QI-group* [25].

We 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 $\mathcal{MM}$ contains k or more tuples.

Based on this definition, in a masked microdata that satisfies k-anonymity property, the probability to correctly identify an individual is at most $1/k$. By increasing k the level of protection increases, along with the changes to the initial microdata.

Unfortunately, k-anonymity does not provide the amount of confidentiality required for every individual ([13], [20], [24]). 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 [20]. This attack is called homogeneity attack [13] and can be avoided by enforcing a more powerful anonymity model than k-anonymity, for example p-sensitive k-anonymity.

Table 3: Masked microdata example for p-sensitive k-anonymity property

Age	ZipCode	Diagnosis	Income
20	41099	AIDS	60,000
20	41099	AIDS	60,000
20	41099	AIDS	40,000
30	41099	Diabetes	50,000
30	41099	Diabetes	40,000
30	41099	Tuberculosis	50,000
30	41099	Tuberculosis	40,000

Definition 3 ***(p-sensitive k-anonymity property)****:* A 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 $\mathcal{MM}$.

To illustrate this property, we consider the masked microdata from Table 3 where *Age* and *ZipCode* are quasi-identifier attributes, and *Diagnosis* and *Income* are confidential attributes:

The above masked microdata satisfies 3-anonymity property with respect to *Age* and *ZipCode*. To determine the value of p, we analyze each QI-cluster with respect to their confidential attribute values. The first QI-cluster (the first three tuples in Table 3 has two different incomes (*60,000* and *40,000*), and only one diagnosis (*AIDS*), therefore the highest value of p for which p-sensitive 3-anonymity holds is 1. As a result, a presumptive intruder who searches information about a young person in his twenties that lives in zip code area 41099 will discover that the target entity suffers from *AIDS*, even if he doesn't know which tuple in the first QI-cluster corresponds to that person. This attribute disclosure problem can be avoided if one of the tuples from the first QI-cluster would have a value other than *AIDS* for *Diagnosis* attribute. In this case, both QI-clusters would have two different illnesses and two different incomes, and, as a result, the highest value of p would be 2.

3. Data Generalization

As illustrated by the previous examples, a general method widely used for masking initial microdata to conform to an anonymity model is the generalization of the quasi-identifier attributes. Other methods that can be used for data anonymization are tuple suppression, data swapping, sampling, etc. They are rarely used and are not discussed in this paper. We define next what generalization is and we describe different existing generalization types.

Generalization of a quasi-identifier attribute consists of replacing the actual value of the attribute with a less specific, more general value that is faithful to the original [19].

Initially, this technique was used for *categorical* attributes and employed predefined (static) domain and value generalization hierarchies [19]. Generalization was extended for *numerical* attributes either by using *predefined hierarchies* [8] or a *hierarchy-free model* [10]. To each categorical attribute a *domain generalization hierarchy* is associated. The values from different domains of this hierarchy are represented in a tree called *value generalization hierarchy*. We illustrate domain and value generalization hierarchy in Figure 1 for attributes *ZipCode* and *Gender*.

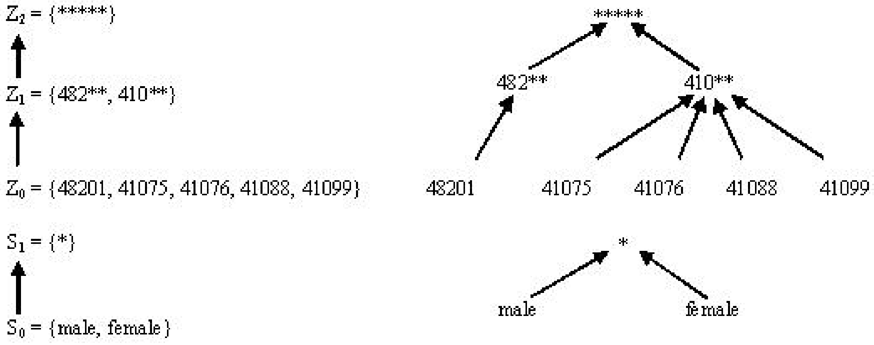

Figure 1: Examples of domain and value generalization hierarchies

There are several ways to perform generalization. Generalization that maps all values of a quasi-identifier categorical attribute to a more general domain in its domain generalization hierarchy is called *full-domain generalization* ([17], [10]). Generalization can also map attribute values to different domains in its domain generalization hierarchy, each value being replaced by the same generalized value in the entire dataset [8]. These two generalization models are also known in statistical literature as global recoding [26]. The least restrictive generalization, called *cell level generalization* [12], extends the second model by allowing the same value to be mapped to different generalized values, in distinct tuples. This generalization method is also known as local recoding

Table 4: Examples of different types of generalizations

Tuples	Age	ZipCode	Gender
r_1	25	41076	Male
r_2	25	41075	Male
r_3	35	41099	Female
r_4	38	48201	Female
r_5	36	41075	Female

$\mathcal{IM}$, k=2, *ZipCode* and *Gender* are categorical attributes with the hierarchies defined in Figure 1

Tuples	Age	ZipCode	Gender
r_1	20-30	*****	Male
r_2	20-30	*****	Male
r_3	30-40	*****	Female
r_4	30-40	*****	Female
r_5	30-40	*****	Female

$\mathcal{MM}$, k=2, full-domain generalization (Iyengar generalization is identical in this case)

Tuples	Age	ZipCode	Gender
r_1	20-30	410**	Male
r_2	20-30	410**	Male
r_3	30-40	*****	Female
r_4	30-40	*****	Female
r_5	30-40	*****	Female

$\mathcal{MM}$, k=2, cell-level generalization

[26]. We illustrate in Table 4 the differences between the above-mentioned types of generalization.

Generalization of numerical attributes using predefined hierarchies is similar to the generalization for categorical attributes. The hierarchy-free generalization replaces the set of values to be generalized to the smallest interval that includes all the initial values. For instance, the values: 35, 38, 36 for the attribute Age are generalized to the interval [35-38]. Note that overlapping of the intervals formed during generalization is possible.

4. Privacy Algorithms

From now on, as k-anonymity is a less restrictive model than p-sensitive k-anonymization, and, correspondingly, transforming a dataset to conform to k-anonymity is easier than transforming it to conform to p-sensitive k-anonymity, we will refer to the p-sensitive k-anonymization problem. K-anonymization is a simpler problem, less complex subclass of p-sensitive k-anonymization. In this section, we will first define the data anonymization problem mentioned above, and, then, we will shortly present three algorithms that find solutions for this problem.

4.1. Data Anonymization Problem Definition

The microdata p-sensitive k-anonymization problem can be formulated as follows:

Definition 4 ***(p-sensitive k-anonymization problem)****:* Given a microdata $\mathcal{IM}$, the p-sensitive k-anonymization problem for $\mathcal{IM}$ is to find a partition $\mathcal{S} = \{cl_1, cl_2, \ldots, cl_v\}$ of $\mathcal{IM}$, where $cl_j \subseteq \mathcal{IM}, j = 1..v$, are called clusters and: $\cup_{j=1}^{v} cl_j = \mathcal{IM}$; $cl_i \cap cl_j = \emptyset, i, j = 1..v, i \neq j$; $|cl_j| \geq k$ and cl_j is p-sensitive, $j = 1..v$; and a cost measure is optimized.

4.2. Local Recoding Anonymization Algorithms

Anonymization algorithms, in addition to achieving the properties required by the target privacy model (in this case p-sensitive k-anonymity, otherwise k-anony-mity, l-diversity, (α, k)-anonymity, t-closeness), must also consider minimizing one or more cost measures. We know that optimal k-anonymization is a NP-hard problem [3]. By simple reduction to k-anonymity, it can be easily shown that p-sensitive k-anonymization is also a NP-hard problem. Several polynomial algorithms that achieve a suboptimal solution currently exist for enforcing p-sensitive k-anonymity and other similar models on microdata. We will refer to three of them.

In [5] we described a greedy clustering algorithm, called *GreedyPKClustering*, for p-sensitive k-anonymity. [22] presents another anonymization algorithm, called *EnhancedPKClustering*, that takes advantage of the known properties of the p-sensitive k-anonymity model in order to improve the p-sensitive k-anonymous solutions w.r.t. various cost measures.

Both these algorithms follow the following "recipe". 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 quasi-identifier attributes generalization. For categorical attributes we used generalization based on predefined hierarchies [8], while for numerical attributes we used the hierarchy-free generalization [10].

The key element in both algorithms is the clusters formation step. We explain briefly the logic behind this step in both *GreedyPKClustering* and *EnhancedPKClustering*.

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 in *GreedyPKClustering* is to ensure that each cluster has enough diversity, i.e. enough distinct values, 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 information loss 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 *GreedyPKClustering* algorithm uses two measures: one for cluster diversity and one for information loss, which correspond to the two criteria explained above. We will introduce later (Subsection 4.4) some of the cost measures used by the anonymization algorithms, which also serve to compare the quality of the masked microdata they produce - among these measures, it will be the information loss measure we mentioned.

The *EnhancedPKClustering* algorithm relies on the cluster formation process on a maximal property that is valid for the p-sensitive k-anonymity. More specifically, given an initial dataset, and a value for p, one can express a superior limit of the number of p-sensitive QI-clusters that can be formed in that microdata set, based on the distribution of sensitive attributes' values. The *EnhancedPKClustering* algorithm places uttermost attention on the p-sensitive part of the model, and starts by enforcing it using the properties proved for the p-sensitive k-anonymity model. The tuples from $\mathcal{IM}$ are distributed to form p-sensitive clusters with respect to the sensitive attributes. This choice has a logical explanation: while k-anonymity is satisfied for each individual cluster when its size is k or more, the p-sensitive property is not so obvious to achieve. After p-sensitivity is achieved, the clusters are further processed to satisfy the k-anonymity requirement as well.

4.3. Global Recoding Anonymization Algorithms

An algorithm proposed initially for k-anonymization of a microdata set is *Incognito* [9]. Subsequently, adapted versions of *Incognito* were proposed to be used when enforcing microdata to conform to other anonymity models, such as l-diversity and (α, k)-anonymity. *Incognito* can be adapted for p-sensitive k-anonymity as well, and we use such an adapted version of *Incognito* in our experiments.

Incognito is a global recoding algorithm, i.e. it generalizes all the values for a quasi-identifier attribute in $\mathcal{IM}$ to ancestors that are at the same level in that attribute value hierarchy. To choose the generalization hierarchy level that will be used for each of the quasi-identifiers attributes when masking the microdata, *Incognito* proceeds as follows. It selects, from all possible generalization level combinations for the quasi-identifier attributes, those that produce k-anonymous masked datasets and are minimal, i.e. there are no less-general level combinations that also would produce k-anonymous masked microdata. For example, *Incognito* will select $< Z_1, S_1 >$ to be a solution for the k-anonymization problem of a microdata set $\mathcal{IM}$ if generalizing the *Sex* and *Zipcode* attributes values in $\mathcal{IM}$ one level each will produce a k-anonymous microdata set $\mathcal{MM}$ and none of $< Z_0, S_0 >$, $< Z_0, S_1 >$ or $< Z_1, S_0 >$ would do the same. (the hierarchies for *Zipcode* and *Sex* are those in Figure 1). *Incognito* will provide more than one solution to the k-anonymization problem; it will find, in fact, all the existing minimal (in the sense described above) solutions of the k-anonymization problem, for a given microdata set $\mathcal{IM}$ and a given k.

The search space where *Incognito* can find the solutions to the k-anonymization problem, for a given microdata set $\mathcal{IM}$ and a given k, is the Cartesian product of domain generalization hierarchy levels for all the quasi-identifier attributes in $\mathcal{IM}$. *Incognito* will not exhaustively explore this search space. Instead, based on some anti-monotone properties of the search space elements w.r.t. the k-anonymity model, *Incognito* prunes the search space and significantly reduces the search for solutions.

Incognito was easily adapted to solve the p-sensitive k-anonymization problem. The search space remains the same, and only the condition checked against a search space element to decide if it represents a solution or not was changed. Namely, a search space element will be a solution if it respects the p-sensitive requirement as well, besides the k-anonymity and minimal requirements that decided the solutions in the original *Incognito* algorithm.

4.4 Cost Measures Used by Anonymization Algorithms

We detail next some of the cost measures that can be used as optimization criteria for the p-sensitive k-anonymization problem ([3], [4], etc.); some of these measures will also serve for comparing the quality of the solutions provided by the three alternate anonymization algorithms in the next section.

A simple cost measure is based on the size of each cluster from a solution $\mathcal{S}$ of the p-sensitive k-anonymization problem. This measure, called *discernability metric* (DM) [3] assigns to each record x from $\mathcal{IM}$ a penalty that is determined by the size of the cluster containing x:

$DM(\mathcal{S}) = \sum_{j=1}^{v} (|cl_j|)^2$.

LeFevre introduced an alternative measure, called the *normalized average cluster size metric* (AVG) [10]:

$AVG(\mathcal{S}) = \frac{n}{v \cdot k}$,

where n is the size of the $\mathcal{IM}$, v is the number of clusters, and k is as in k-anonymity. We notice that the AVG cost measure is inversely proportional to the number of clusters.

The last cost measure we describe is the *information loss* caused by generalizing each cluster to a common tuple ([4], [21]). This is an obvious measure that guided the partitioning process in both *GreedyPKClustering* and *EnhancedPKClustering*, since the partition $\mathcal{S}$ these algorithms find as solution to the p-sensitive k-anonymity problem is subsequently subject to cluster-level generalization.

To introduce information loss measure, we have to define first the *generalization information* for a cluster in $\mathcal{S}$. We call generalization information for a cluster the minimal covering tuple for that cluster, and we define it as follows.

Definition 5 ***(generalization information):*** Let $cl = \{r_1, r_2, \ldots, r_q\} \in \mathcal{S}$ be a cluster, $\mathcal{KN} = \{N_1, N_2, \ldots, N_s\}$ be the set of numerical quasi-identifier attributes and $\mathcal{KC} = \{C_1, C_2, \ldots, C_t\}$ be the set of categorical quasi-identifier attributes. The *generalization information* of *cl*, w.r.t. quasi-identifier attribute set $\mathcal{K} = \mathcal{KN} \cup \mathcal{KC}$ is the "tuple" *gen*(*cl*), having the scheme K, where:

- For each categorical attribute $C_j \in \mathcal{K}$, $gen(cl)[C_j]$ = the lowest common ancestor in H_{C_j} of $\{r_1[C_j], r_2[C_j], \ldots, r_q[C_j]\}$, where H_C denotes the hierarchies (domain and value) associated to the categorical quasi-identifier attribute C;
- For each numerical attribute $N_j \in \mathcal{K}$, $gen(cl)[N_j]$ = the interval $[min\{r_1[N_j], r_2[N_j], \ldots, r_q[N_j]\}, max\{r_1[N_j], r_2[N_j], \ldots, r_q[N_j]\}]$.

For a 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 $\mathcal{MM}$, each tuple from the cluster cl will be replaced by $gen(cl)$.

Definition 6 ***(cluster information loss)**:* Let $cl \in \mathcal{S}$ be a cluster, $gen(cl)$ its generalization information and $\mathcal{K} = \{N_1, N_2, \ldots, N_s, C_1, C_2, \ldots, C_t\}$ the set of quasi-identifier attributes. The *cluster information loss* caused by generalizing cl tuples to $gen(cl)$ is:

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

where:

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

Definition 7 ***(total information loss)**:* Total information loss for a solution $\mathcal{S} = \{cl_1, cl_2, \ldots, cl_v\}$ of the p-sensitive k-anonymization problem, denoted by $IL(\mathcal{S})$, is the sum of the information loss measure for all the clusters in $\mathcal{S}$:

$$IL(\mathcal{S}) = \sum_{j=1}^{v} IL(cl_j).$$

5. Experimental Results

In this section we compare the performance of the *GreedyPKClustering* algorithm, the *EnhancedPKClustering* algorithm and the adapted version of the *Incognito* algorithm. 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, (α, k)-anonymity, and t-closeness). However, as supported by the current experiments, we think that an algorithm based on global recoding (such as *Incognito*) will generally produce weaker results, in terms of any cost measure, compared to a local recoding algorithm (such as *EnhancedPKClustering* or *GreedyPKClustering*). This is without connection to a specific anonymity model.

All three algorithms have been implemented in Java, and tests were executed on a dual CPU machine running Windows 2003 Server with 3.00 GHz and 1 GB of RAM.

Table 5: Data distribution in the synthetic datasets

	All QI Attributes	**All Sensitive Attributes**
$Dataset\_UU$	Uniform	Uniform
$Dataset\_UN$	Uniform	Normal
$Dataset\_NU$	Normal	Uniform
$Dataset\_NN$	Normal	Normal

Table 6: Mapping between 0-8 range and discrete values

$val < 1$	$1 \leq val < 2$	$2 \leq val < 3$	...	$6 \leq val < 7$	$7 \leq val$
a	b	c	...	g	h

A set of experiments has been conducted for an $\mathcal{IM}$ consisting of 10000 tuples randomly selected from the *Adult* dataset from the UC Irvine Machine Learning Repository [14]. 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 with respect to the quasi-identifier consisting of all 6 quasi-identifier attributes and all 3 confidential attributes.

Another set of experiments used synthetic datasets, where the quasi-identifier and the sensitive attributes' values were generated to follow some predefined distributions. For our experiments, we generated four microdata sets using normal and uniform distribution. All four data sets have identical schema $(QI\_N, QI\_C1,\ QI\_C2, QI\_C3, S\_C1, S\_C2)$ where the first attribute (QI_N) is quasi-identifier of type numerical (*Age* like), the next three $(QI\_C1, QI\_C2,$ $QI\_C3)$ are categorical quasi-identifiers and the last two (S_C1 and S_C2) are categorical sensitive attributes. The distribution followed by each attribute for the four data sets are illustrated in Table 5.

For the numerical attribute we use the *age* like values 0, 1, ..., 99. To generate a uniform distribution for this range we use the mean 99/2 and standard deviation of 99/6. For each categorical attribute we use 8 values that are grouped in a hierarchy as shown in Figure 2. To generate a uniform-like distribution for the categorical attributes we use the range 0-8 with mean 8/2 and standard deviation 8/6 and the mapping shown in Table 6 (*val* is the value computed by the generator).

We present next the experimental results we obtained, which prove the

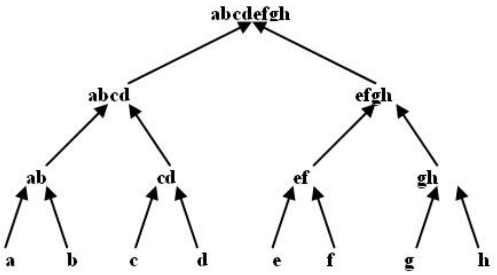

Figure 2: The value generalization hierarchy for the categorical attributes of the synthetic datasets

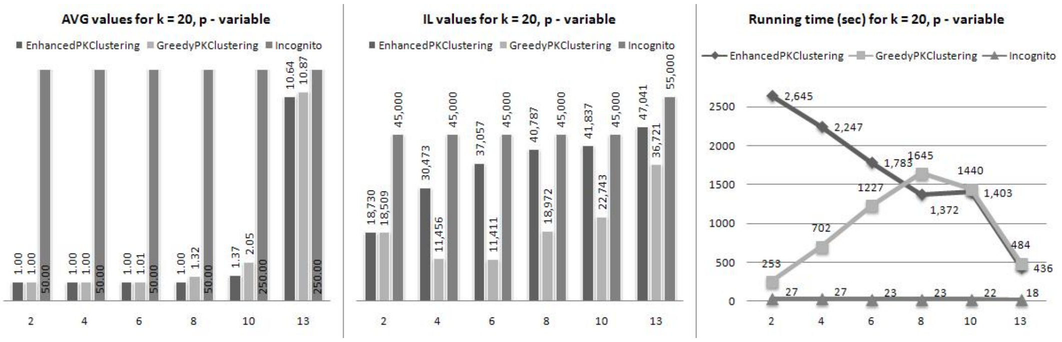

Figure 3: *AVG*, *IL*, *RT* for the 3 algorithms, *Adult* dataset

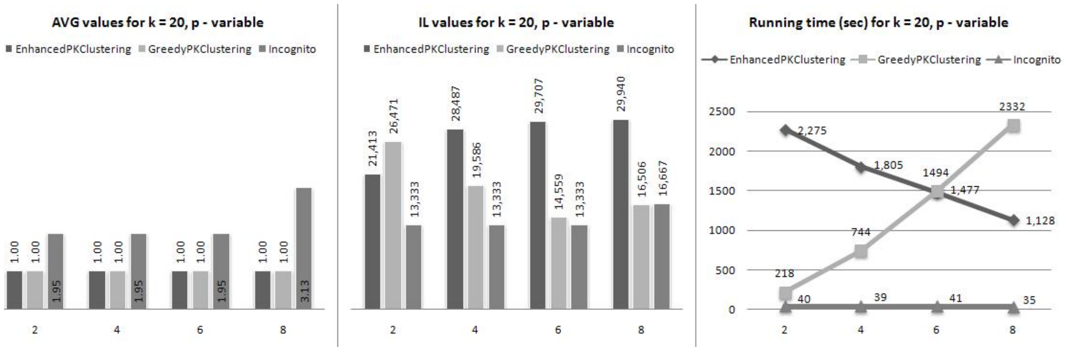

Figure 4: *AVG*, *IL*, *RT* for the 3 algorithms, *Dataset_UU*

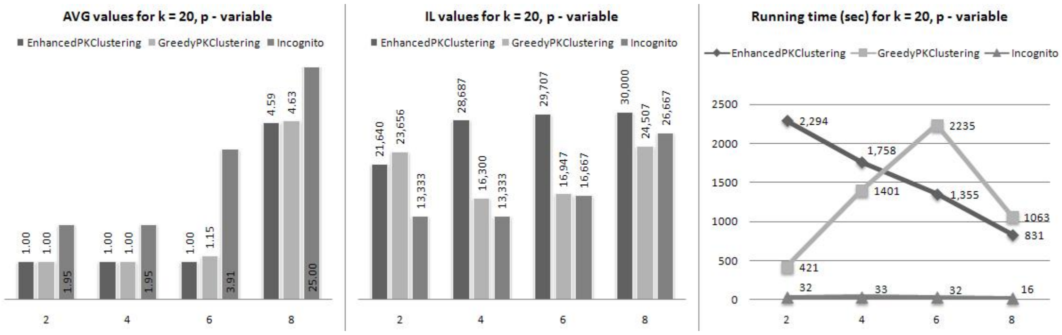

Figure 5: *AVG*, *IL*, *RT* for the 3 algorithms, *Dataset_UN*

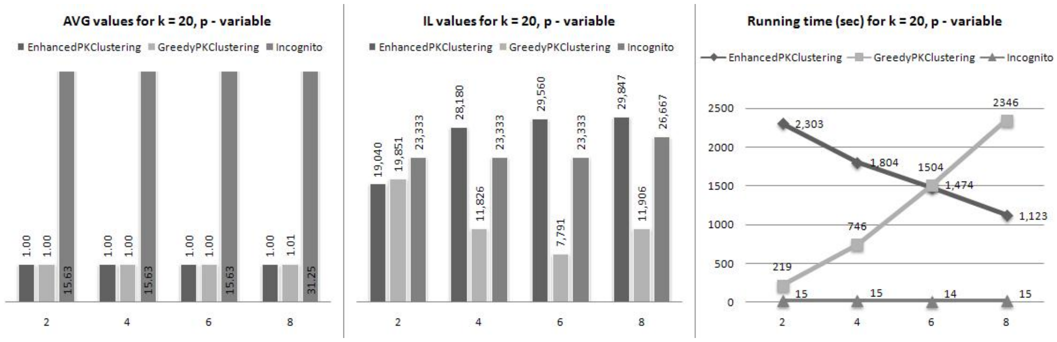

Figure 6: *AVG*, *IL*, *RT* for the 3 algorithms, *Dataset_NU*

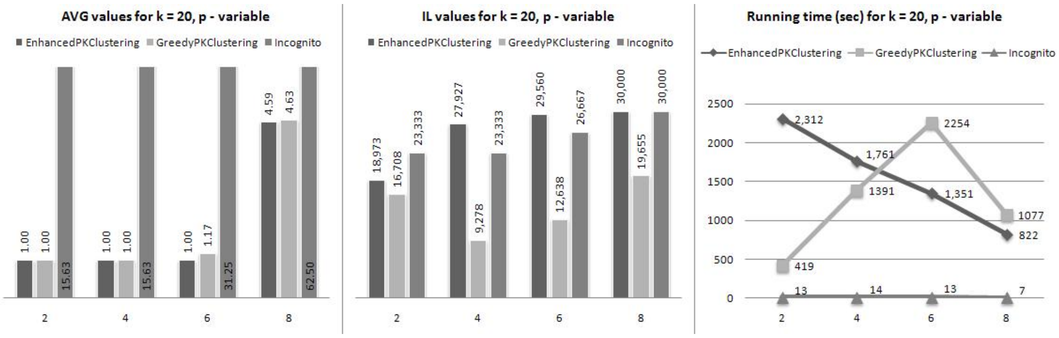

Figure 7: *AVG*, *IL*, *RT* for the 3 algorithms, *Dataset_NN*

above statements, for the p-sensitive k-anonymization problem. However, these facts should stand for the k-anonymization problem as well, as this problem is a simpler case of the more complex p-sensitive k-anomymization problem. Concretely, we present, for each of the five experimental datasets we used, the *AVG*, *IL*, and execution time cost measure values, for each of the three algorithms, *GreedyPKClustering*, *EnhancedPKClustering* and *Incognito*, for different k and p values.

Figures 3 to 7 show comparatively the *AVG*, the *IL*, and the running time values of the three algorithms, *EnhancedPKClustering*, *GreedyPKClustering* and *Incognito*, produced for $k = 20$ and different p values, for the *Adult* dataset (Figure 3), Dataset_UU (Figure 4), Dataset_UN (Figure 5), Dataset_NU (Figure 6), and Dataset_NN (Figure 7).

The *AVG* results for the first two algorithms clearly outperform *Incognito* in all experiments. We notice that *EnhancedPKClustering* is able to improve in many cases the performances of the *GreedyPKClustering* algorithm in cases where solving the p-sensitivity part takes prevalence over creating clusters of size k. We notice that for $p = 2$ and 4 there is no improvement. In these cases both algorithms were able to find the optimal solution in terms of *AVG* values. As soon as the p-sensitive part is difficult to achieve, the *EnhancedPKClustering* algorithm performs better.

The *IL* results are strongly related to the quasi-identifier attributes distribution, for the *GreedyPKClustering* and *Incognito* algorithms. Surprisingly, they work in an opposite manner to each other: *Incognito* performs better when quasi-identifier attributes follow a uniform distribution, while *GreedyPKClustering* obtains better results when quasi-identifier attributes follow a normal distribution. *EnhancedPKClustering* seems to be independent of the quasi-identifiers distribution. Overall, *GreedyPKClustering* seems to obtain the best *IL* results.

W.r.t. the time required to generate the masked microdata by the compared algorithms, we notice the following. Since *Incognito* uses global recording and our domain generalization hierarchies for this datasets have a low height, the running time is very fast. The *GreedyPKClustering* is faster than the *EnhancedPKClustering* algorithm for small values of p, but when it becomes more difficult to create p-sensitivity within each cluster the *EnhancedPKClustering* has a slight advantage. We also notice that the running time of the *GreedyPKClustering* algorithm is influenced by the sensitive attributes' distribution.

6. Conclusions

In this paper we presented an extensive overview and comparison between two local recoding anonymization algorithms (*GreedyPKClustering*, *EnhancedPKClustering*), and a global recoding anonymization algorithm (*Incognito*) in terms of *AVG*, *IL*, and running time. Our experiments have shown that the local recoding algorithms outperform *Incognito* in terms of *AVG* measure. While it is a lot harder to draw a similar conclusion for *IL* measure, we notice that the quasi-identifier attributes distribution is an important component in the *IL* results. All three algorithms obtain similar *IL* results, with *GreedyPKClustering* being a close winner. As expected, the running time of *Incognito* is lower than the running time of both local recoding algorithms (as the search space of *Incognito* is a lot less complex than for the others, which explore the microdata tuples space). Based on our experiments, we think that a local recoding algorithm should be preferred in order to obtain good quality results, when the running time is not critical.

References

[1] Aggarwal G., Feder T., Kenthapadi K., Khuller S., Panigrahy R., Thomas D., Zhu A., Anonymizing Tables, Proc. of the ACM PODS Conference, (2006), 153-162.

[2] Agrawal R., Kiernan J., Srikant R., Xu Y., Hippocratic Databases, Proc. of the 20th International Conference on Very Large Databases (VLDB), Hong Kong, (2002), 143-154.

[3] Bayardo R.J, Agrawal R., Data Privacy through Optimal k-Anonymization, Proc. of the IEEE International Conference of Data Engineering, (2005), 217-228.

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

[5] Campan A., Truta T.M., Miller J., Sinca R., A Clustering Approach for Achieving Data Privacy, The 2007 International Conference on Data Mining (DMIN2007), Las Vegas, (2007), 321-327.

[6] Gramm-Leach-Bliley Financial Services Modernization Act, Available online at http://banking.senate.gov/conf/, (1999).

[7] Health Insurance Portability and Accountability Act, Available online at http://www.hhs.gov/ocr/hipaa, (2002).

[8] Iyengar V., Transforming Data to Satisfy Privacy Constraints, Proc. of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (2002), 279-288.

[9] LeFevre K., DeWitt D., Ramakrishnan R., Incognito: Efficient Full-Domain K-Anonymity, Proc. of the ACM SIGMOD, Baltimore, Maryland, (2005), 49-60.

[10] LeFevre K., DeWitt D., Ramakrishnan R., Mondrian Multidimensional K-Anonymity, Proc. of the IEEE International Conference of Data Engineering, Atlanta, Georgia, (2006).

[11] Li N., Li T., Venkatasubramanian S., T-Closeness: Privacy Beyond k-Anonymity and l-Diversity, Proc. of the IEEE ICDE (2007).

[12] Lunacek M., Whitley D, Ray I., A Crossover Operator for the k-Anonymity Problem, Proc. of the GECCO Conference, (2006), 1713-1720.

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

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

[15] Privacy and Human Rights: An International Survey of Privacy Laws and Developments, available online at http://www.privacyinternational.org/survey/phr2003/countries/romania.htm, (2003).

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

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

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

[19] Sweeney L., Achieving *k*-Anonymity Privacy Protection Using Generalization and Suppression, International Journal on Uncertainty, Fuzziness, and Knowledge-based Systems, Vol. 10, No. 5, (2002), 571-588.

[20] Truta T.M., Bindu V., Privacy Protection: P-Sensitive K-Anonymity Property, Proc. of the Workshop on Privacy Data Management, In Conjunction with 22th IEEE International Conference of Data Engineering (ICDE), Atlanta, Georgia, (2006), ?-?.

[21] Truta T.M., Campan A., K-Anonymization Incremental Maintenance and Optimization Techniques, Proc. of the ACM SAC, (2007), 380-387.

[22] Truta T.M., Campan A., Meyer P., Generating Microdata with P-Sensitive K-Anonymity Property, accepted to the 4th VLDB Workshop on Secure Data Management (SDM2007), Vienna, Austria, (2007).

[23] Winkler W.E., Advanced Methods for Record Linkage, Proc. of the Section on Survey Research Methods, American Statistical Society, (1994), 467-472.

[24] 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), 754-759.

[25] Xiao X., Tao Y., Personalized Privacy Preservation, Proc. of the ACM SIGMOD, (2006), 229-240.

[26] Xu J., Wang W., Pei J., Wang X., Shi B., Fu A. W.-C., Utility-based anonymization using local recoding, Proc. of the 12th ACM SIGKDD international conference on knowledge discovery and data mining, (2006), 785-790.

Authors:

Traian Marius Truta, Michael Abrinica, John Miller
Department of Computer Science

Northern Kentucky University
Highland Heights, KY 41099, USA
emails:*trutat1@nku.edu*, *abrinicam@nku.edu*, *millerj10@nku.edu*

Alina Campan
Department of Computer Science
Babes-Bolyai University of Cluj-Napoca
Str. Mihail Kogalniceanu nr. 1, RO-400084, Cluj-Napoca
email:*alina@cs.ubbcluj.ro*