

# P-Sensitive K-Anonymity with Generalization Constraints

Alina Campan\*, Traian Marius Truta\*, Nicholas Cooper\*

\* Department of Computer Science

Northern Kentucky University

Highland Heights, KY 41099, USA

E-mail: {campana1, trutat1, coopern1}@nku.edu

**Abstract.** Numerous privacy models based on the  $k$ -anonymity property and extending the  $k$ -anonymity model have been introduced in the last few years in data privacy research:  $l$ -diversity,  $p$ -sensitive  $k$ -anonymity,  $(\alpha, k)$ -anonymity,  $t$ -closeness, etc. While differing in their methods and quality of their results, they all focus first on masking the data, and then protecting the quality of the data as a whole. We consider a new approach, where requirements on the amount of distortion allowed on the initial data are imposed in order to preserve its usefulness. Our approach consists of specifying quasi-identifiers' generalization constraints, and achieving  $p$ -sensitive  $k$ -anonymity within the imposed constraints. We think that limiting the amount of allowed generalization when masking microdata is indispensable for real life datasets and applications. In this paper, the *constrained  $p$ -sensitive  $k$ -anonymity model* is introduced and an algorithm for generating constrained  $p$ -sensitive  $k$ -anonymous microdata is presented. Our experiments have shown that the proposed algorithm is comparable with existing algorithms used for generating  $p$ -sensitive  $k$ -anonymity with respect to the results' quality, and obviously the obtained masked microdata complies with the generalization constraints as indicated by the user.

## 1 Introduction

Large amounts of personal data are constantly being collected in various application fields: for taxes, healthcare, in credit card transactions, in social networks etc. Besides its primary purpose for which it is collected in the first place, data is subsequently analyzed and mined by its owners and/or third parties before being retired. But as stipulated by existing regulations in various countries and for different application domains, the privacy of the individuals described in electronic datasets must be protected during their collection, storage, distribution, and use [12].

---

Therefore, before publishing or releasing a microdata set (a dataset where each tuple corresponds to one individual), a frequently used solution is to modify the initial microdata, in order to enforce an anonymity model on it. This modification (also known as masking) is being performed to counter record linkage attacks [34] between the masked dataset and external available information, based on a set of attributes called quasi-identifier or key attributes; if successful, such attacks would allow gleaning the identity of individuals or their private information from the released microdata.

Among the anonymity models targeted by the masking process are  $k$ -anonymity [26, 27] or one of its extensions:  $l$ -diversity [21],  $p$ -sensitive  $k$ -anonymity [29],  $(\alpha, k)$ -anonymity [35],  $t$ -closeness [16],  $(k, e)$ -anonymity [37],  $(c, k)$ -safety [22],  $m$ -confidentiality [35], personalized privacy [31],  $m$ -invariance [32],  $\delta$ -presence [24], skyline privacy [6],  $(\epsilon, m)$ -anonymity [17],  $l^+$ -diversity [18],  $(\tau, \lambda)$ -uniqueness [33],  $(k, p, q, r)$ -anonymity [7] etc. A microdata set conforms to the  $k$ -anonymity property if every tuple within it is indistinguishable of at least  $(k-1)$  other tuples, with respect to the set of quasi-identifier attributes. The other mentioned models require extra conditions compared to  $k$ -anonymity, for instance, that a certain count or distribution of sensitive attribute values are ensured for every group of  $k$  or more tuples that share common quasi-identifier values.

Irrespective of the targeted anonymity model and the methods employed to achieve it, two contradicting goals are generally followed in the masking process: creating an anonymous dataset while preserving as much as possible the informational content of the initial dataset. Most of the existing work focuses first on achieving the anonymity model without limiting the amount of allowed generalization. Therefore, even if the masked microdata is of good quality and preserves well the overall content in the initial microdata, it can still be useless because necessary information has been lost for critical attributes of the data. There are several recent research papers that address this issue. Short description of these papers and how they differ from this paper are presented in the related work section (Section 5).

We introduce in the next section the *constrained  $p$ -sensitive  $k$ -anonymity model*, that protects against identity and attribute disclosure [13], while keeping the quasi-identifiers' generalization restricted to certain user-specified constraints (or boundaries). We also describe in Section 3 an algorithm for transforming a microdata set to conform to our new anonymity model. Experimental results are reported in Section 4 for  $p$ -sensitive  $k$ -anonymity, constrained and not, that illustrate the quality of masked microdata and the masking algorithms' efficiency. Related work is presented in Section 5. Conclusions are given in Section 6.

## 2 Anonymity and Constraints

An initial microdata  $IM$  is a set of tuples over a set of attributes. The attributes describing the initial microdata set are usually categorized in the following three types:

- *Identifier* attributes such as *Name* and *SSN* that can be used to identify a record. These attributes are removed as part of the masking process because they express information which can directly lead to a specific entity. We will denote the identifier attributes by  $I_1, I_2, \dots, I_m$ .
- *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. However, quasi-identifier attribute values are possibly altered in the masked microdata in order to prevent identity and attribute disclosure. We will denote the quasi-identifiers by  $Q_1, Q_2, \dots, Q_t$ .
- *Sensitive* or *confidential* 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, and are usually kept unmodified. We will denote the sensitive attributes by  $S_1, S_2, \dots, S_r$ .

Let  $IM$  be a microdata with the schema  $\mathcal{R} = \{ I_1, I_2, \dots, I_m, Q_1, Q_2, \dots, Q_t, S_1, S_2, \dots, S_r \}$ . We define next what its corresponding constrained  $p$ -sensitive  $k$ -anonymous masked microdata  $\mathcal{MM}$  is. First of all,  $\mathcal{MM}$  will have the schema  $\mathcal{R}' = \{ Q_1, Q_2, \dots, Q_t, S_1, S_2, \dots, S_r \}$ , and at most  $|IM|$  tuples; it will satisfy the  $k$ -anonymity and  $p$ -sensitivity properties, and it will be free of constraint violations – all these concepts are defined and explained next.

**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.

**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.

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

Based on these definitions, in a masked microdata that satisfies  $p$ -sensitive  $k$ -anonymity, the probability to correctly identify an individual is at most  $1/k$  and the probability that some sensitive information about the individual is disclosed

is at most  $1/p$ . By increasing  $k$  and  $p$  the level of protection increases, usually along with the changes needed to anonymize the initial microdata.

As mentioned before, the initial microdata undergoes an anonymization process that transforms it to comply with the  $p$ -sensitive  $k$ -anonymity model. The techniques most often employed in any anonymization process are the generalization of quasi-identifier attribute values and tuple suppression [28]. The generalization of a quasi-identifier attribute consists in replacing the actual values of the attribute with less specific, more general values that are faithful to the original [28]. In this paper we use generalization based on predefined generalization hierarchies [14], for both numerical and nominal attributes. Suppressing entire tuples allows sometimes reducing the amount of generalization required for achieving  $p$ -sensitive  $k$ -anonymity.

Anonymization algorithms generally try to achieve the desired protection level with minimal changes (as quantified by an information loss measure [4]) to the initial microdata  $IM$ . However, minimal changes may correspond to generalization that surpasses a data usefulness threshold, beyond which the masked microdata  $MM$  would become unusable. Imagine a healthcare microdata that includes address information of the patients and of the medical services' providers. If a study intends to verify a hypothesis that patients travel mostly locally (county or state level) for medical services, but the available masked microdata contains geographic information generalized to the country level, then the study cannot be performed, due to inappropriate granularity of the *Location* values. In this example the *Location* values should not be generalized beyond the state level; in other words, the states should constitute generalization boundaries for this attribute's values.

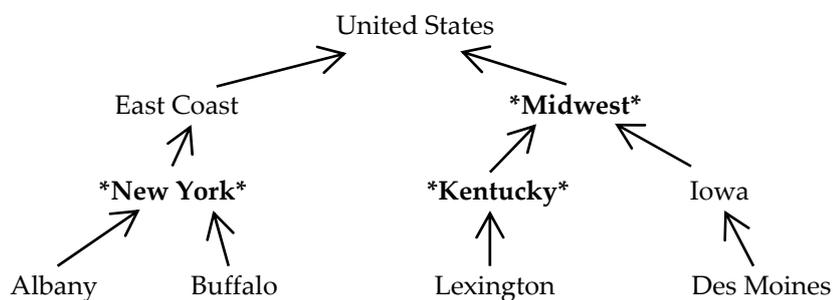
To allow expressing generalization boundaries, we associate with each quasi-identifier attribute value a *maximum allowed generalization value*. This concept is used to express how far the owner of the data thinks that the quasi-identifier's values could be generalized, without causing the resulting masked microdata to be useless. The data owner is often aware of the way various research studies are using the data and he/she is able to identify maximum allowed generalization values.

**Definition 4 (Maximum Allowed Generalization Value).** Let  $Q$  be a quasi-identifier attribute, and  $\mathcal{H}_Q$  its predefined value generalization hierarchy. For every leaf value  $\mu \in \mathcal{H}_Q$ , the *maximum allowed generalization value* of  $\mu$ ,  $MAGVal(\mu)$ , is the value (leaf or not-leaf) in  $\mathcal{H}_Q$  situated on the path from  $\mu$  to the root, such that for any released microdata, the value  $\mu$  is permitted to be generalized only up to  $MAGVal(\mu)$ .

Figure 1 contains an example of defining maximal allowed generalization values for a subset of values for the *Location* attribute.

The *MAGVals* for the leaf values “Albany” and “Des Moines” are “New York”, and, respectively, “Midwest” (the maximal allowed generalization values are marked with \* characters that delimit them). This means that the quasi-identifier *Location*’s value “Albany” may be generalized to itself or to “New York”, but not to “East Coast” nor the “United States”. Also, “Des Moines” may be generalized to itself, “Iowa”, or “Midwest”, but it may not be generalized to the “United States”.

Figure 1. Example of *MAGVals*.



When several *MAGVals* exist on the path between a particular leaf value,  $\mu$ , and the hierarchy root, then the  $MAGVal(\mu)$  is considered to be the first *MAGVal* that is reached when following the path from  $\mu$  to the root node. (Several such *MAGVals* on a path between a leaf and the root can result from defining *MAGVals* for other leaves within that hierarchy). Therefore, the path between  $\mu$  and  $MAGVal(\mu)$  can contain no node other than  $MAGVal(\mu)$  that is a maximum allowed generalization value.

**Definition 5 (Maximum Allowed Generalization Set).** Let  $Q$  be a quasi-identifier attribute and  $\mathcal{H}_Q$  its predefined value generalization hierarchy. The set of all *MAGVals* for attribute  $Q$  is called  $Q$ ’s **maximum allowed generalization set**, and it is denoted by  $MAGSet(Q) = \{ MAGVal(\mu) \mid \forall \mu \in leaves(\mathcal{H}_Q) \}$  (The notation  $leaves(\mathcal{H}_Q)$  represents all the leaves from the  $\mathcal{H}_Q$  value generalization hierarchy).

Given the hierarchy for the attribute *Location* presented in Figure 1,  $MAGSet(Location) = \{New\ York, Kentucky, Midwest\}$ .

Usually, the data owner/user only has generalization restrictions for some of the quasi-identifiers in a microdata that is to be masked. If for a particular quasi-identifier attribute  $Q$  there are not any restrictions in respect to its

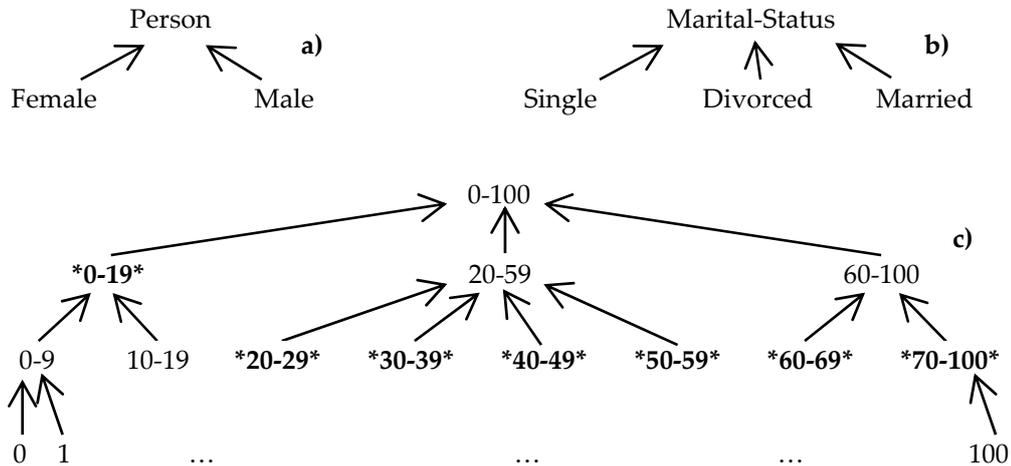
generalization, then no maximal allowed generalization values are specified for  $Q$ 's value hierarchy; in this case, each leaf value in  $\mathcal{H}_Q$  is considered to have the  $\mathcal{H}_Q$ 's root value as its maximal allowed generalization value.

**Definition 6 (Constraint Violation).** We say that the masked microdata  $\mathcal{MM}$  has a *constraint violation* if one quasi-identifier value,  $\mu$ , in  $\mathcal{IM}$ , is generalized in one tuple in  $\mathcal{MM}$  beyond its specific maximal generalization value,  $MAGVal(\mu)$ .

**Definition 7 (Constrained  $P$ -Sensitive  $K$ -Anonymity).** The masked microdata  $\mathcal{MM}$  satisfies the *constrained  $p$ -sensitive  $k$ -anonymity property* if it satisfies  $p$ -sensitive  $k$ -anonymity and it does not have any constraint violation.

Consider the following example. The initial microdata set  $\mathcal{IM}$  in Table 1 is characterized by the following attributes: *Name* is an identifier attribute (to be removed from the masked microdata), *Marital-Status*, *Gender*, and *Age* are the quasi-identifier attributes, and *Diagnosis* is the sensitive attribute. The attribute *Age*'s values and their *MAGVals* are described by Figure 2 c). The remaining quasi-identifier attributes do not have any generalization boundary requirements; their value generalization hierarchies are illustrated in Figure 2 a) and b). This microdata set has to be masked such that the corresponding masked microdata will satisfy constrained 2-sensitive 3-anonymity.

**Figure 2. Hierarchies for the quasi-identifier attributes *Gender* (a), *Marital-Status* (b) and *Age* (c).**



Tables 2, 3, and 4 illustrate three possible masked microdata  $\mathcal{MM}_1$ ,  $\mathcal{MM}_2$ , and  $\mathcal{MM}_3$  for the initial microdata  $\mathcal{IM}$ . The first one,  $\mathcal{MM}_1$ , satisfies 3-anonymity, but it is not  $p$ -sensitive 3-anonymous, for any  $p$ ,  $p \geq 2$ . This is because the second *QI*-

cluster  $(X_2, X_6, X_7)$  has no diversity in the *Diagnosis* attribute. Therefore, an intruder looking for information about a woman in her 30s can find out easily that his/her target has flu.  $\mathcal{MM}_2$  is 2-sensitive 3-anonymous, but it contains several constraint violations with respect to *Age* attribute's maximal allowed generalization. On the other hand, the third microdata set,  $\mathcal{MM}_3$ , satisfies constrained 2-sensitive 3-anonymity: every *QI*-cluster consists of 3 tuples, it has at least 2 different *Diagnosis* values and none of the *Age*'s initial values is generalized beyond its *MAGVal*.

<i>Record</i>	<i>Name</i>	<i>Marital-Status</i>	<i>Gender</i>	<i>Age</i>	<i>Diagnosis</i>
$X_1$	Bob	Married	Male	37	Cancer
$X_2$	Nancy	Married	Female	30	Flu
$X_3$	James	Married	Male	36	HIV
$X_4$	Carol	Single	Female	43	Cancer
$X_5$	William	Single	Male	45	Flu
$X_6$	Heidi	Single	Female	37	Flu
$X_7$	Cindy	Single	Female	32	Flu
$X_8$	Michael	Married	Male	30	Diabetes
$X_9$	David	Divorced	Female	41	Diabetes

**Table 1.** An *IM* dataset

<i>Records</i>	<i>Marital-Status</i>	<i>Gender</i>	<i>Age</i>	<i>Diagnosis</i>
$X_1, X_3, X_8$	Married	Male	30-39	Cancer, HIV, Diabetes
$X_2, X_6, X_7$	Mar.-Status	Female	30-39	Flu, Flu, Flu
$X_4, X_5, X_9$	Mar.-Status	Person	40-49	Cancer, Flu, Diabetes

**Table 2.** Masked microdata  $\mathcal{MM}_1$  for *IM*

<i>Records</i>	<i>Marital-Status</i>	<i>Gender</i>	<i>Age</i>	<i>Diagnosis</i>
$X_1, X_3, X_8$	Married	Male	30-39	Cancer, HIV, Diabetes
$X_4, X_5, X_6$	Single	Person	20-59	Cancer, Flu, Flu
$X_2, X_7, X_9$	Mar.-Status	Female	20-59	Flu, Flu, Diabetes

**Table 3.** Masked microdata  $\mathcal{MM}_2$  for *IM*

---

<i>Records</i>	<i>Marital-Status</i>	<i>Gender</i>	<i>Age</i>	<i>Diagnosis</i>
$X_1, X_2, X_8$	Married	Person	30-39	Cancer, Flu, Diabetes
$X_3, X_6, X_7$	Mar.-Status	Person	30-39	HIV, Flu, Flu
$X_4, X_5, X_9$	Mar.-Status	Person	40-49	Cancer, Flu, Diabetes

---

**Table 4. Masked microdata  $\mathcal{MM}_3$  for  $\mathcal{IM}$**

### 3 Algorithm

We present in this section a necessary condition for a microdata set to be amenable to constrained  $p$ -sensitive  $k$ -anonymity, given the generalization constraints for its quasi-identifier attributes, expressed as *MAGVals* in their corresponding hierarchies, and specific  $p$  and  $k$  values. We provide next an algorithm for masking a microdata set to  $p$ -sensitive  $k$ -anonymity while staying within the bounds imposed by generalization constraints and attempting to minimize the information loss caused by generalization and suppression (the two procedures employed in the anonymization algorithm).

Our approach to constrained  $p$ -sensitive  $k$ -anonymization partially follows an idea found in [1] and [3], which consists in modeling and solving anonymization as a clustering problem. Basically, the algorithm takes an initial microdata set  $\mathcal{IM}$  and establishes a “good” partitioning of it into clusters. The released microdata set  $\mathcal{MM}$  is afterwards formed by generalizing the quasi-identifier attributes’ values of all tuples inside each cluster to the same values (called generalization information for a cluster).

**Definition 8 (Generalization Information).** Let  $cl = \{X_1, X_2, \dots, X_u\}$  be a cluster of tuples selected from  $\mathcal{IM}$ ,  $QI = \{Q_1, Q_2, \dots, Q_t\}$  be the set of quasi-identifier attributes. The *generalization information of  $cl$*  with respect to the quasi-identifier attribute set  $QI$  is the “tuple”  $gen(cl)$ , having the scheme  $QI$ , where for each attribute  $Q_j \in QI$ ,  $gen(cl)[Q_j] =$  the lowest common ancestor in  $\mathcal{H}_{Q_j}$  of  $\{X_1[Q_j], X_2[Q_j], \dots, X_u[Q_j]\}$ .

The generalization information for a cluster  $cl$  is the “tuple” whose value for each quasi-identifier attribute is the most specific common ancestor (in the attribute’s value generalization hierarchy) of all that attribute’s values in all tuples in  $cl$ . In the corresponding  $\mathcal{MM}$ , each tuple from the cluster  $cl$  will have its quasi-identifier attributes values replaced by  $gen(cl)$ .

---

The following is a necessary and sufficient condition for the anonymization of a cluster *not* to create constraint violations.

**Property 1.** Let  $IM$  be a microdata set and  $cl$  a cluster of tuples from  $IM$ . The generalization of the tuples from  $cl$  to  $gen(cl)$  will not create any constraint violation if and only if  $MAGVal(X_i[Q]) = MAGVal(X_j[Q])$ , for any two tuples  $X_i$  and  $X_j$  from  $cl$  and any quasi-identifier attribute  $Q$ .

**Proof.**

“If”: If  $MAGVal(X_i[Q]) = MAGVal(X_j[Q])$  for any two tuples  $X_i$  and  $X_j$  from  $cl$  and any quasi-identifier attribute  $Q$ , then  $MAGVal(X_i[Q])$  is a common ancestor for all elements of the set  $\{X_u[Q] \mid \text{for all } X_u \in cl\}$ . Using Definition 8,  $MAGVal(X_i[Q])$  is equal to or is an ancestor of  $gen(cl)[Q]$ , which means that there are no constraint violations for the generalization of attribute  $Q$ 's values from the cluster  $cl$ .

“Only If”: Assume that there are two tuples  $X_i$  and  $X_j$  within  $cl$  such that  $MAGVal(X_i[Q]) \neq MAGVal(X_j[Q])$ , where  $X_i[Q], X_j[Q] \in leaves(\mathcal{H}_Q)$  ( $leaves(\mathcal{H}_Q)$  represents all the leaves from the  $\mathcal{H}_Q$  value generalization hierarchy). Let  $a$  be a value within  $\mathcal{H}_Q$  that is the first common ancestor for  $MAGVal(X_i[Q])$  and  $MAGVal(X_j[Q])$ . As a result,  $a$  will be different from, and an ancestor for at least one of  $MAGVal(X_i[Q])$  or  $MAGVal(X_j[Q])$ . This is a consequence of the fact that  $MAGVal(X_i[Q]) \neq MAGVal(X_j[Q])$ : a common ancestor of two different nodes  $x$  and  $y$  in a tree is a node which is different from at least one of the nodes  $x$  and  $y$ . Because of this fact, when  $cl$  will be generalized to  $gen(cl)$ ,  $gen(cl)[Q]$  will be  $a$  (or depending on the other tuples in  $cl$ , even an ancestor of  $a$ ) – therefore at least one of the values  $X_i[Q]$  and  $X_j[Q]$  will be generalized further than its maximal allowed generalization value, leading to a constraint violation. // q.e.d.

The logical consequence of Property 1 is that a microdata set  $IM$  cannot always be anonymized using generalization only, given certain generalization constraints, and specific  $p$  and  $k$  values. As shown next, there exists a minimal set of tuples from  $IM$  which must be suppressed so that it is possible to build a constrained  $p$ -sensitive  $k$ -anonymous masked microdata for the remaining tuples. The following definition introduces a microdata concept that will help us express when  $IM$  can be transformed to satisfy constrained  $p$ -sensitive  $k$ -anonymity using generalization only.

**Definition 9 (Maximum Allowed Microdata).** The *maximum allowed microdata* for a microdata  $IM$ ,  $MAM$ , is the masked microdata where every quasi-identifier value,  $\mu$ , in  $IM$  is generalized to  $MAGVal(\mu)$ .

---

**Property 2.** For a given  $IM$ , if its maximum allowed microdata  $MAM$  is not  $p$ -sensitive  $k$ -anonymous, then any masked microdata obtained from  $IM$  by applying generalization only will not satisfy constrained  $p$ -sensitive  $k$ -anonymity.

**Proof.** Assume that  $MAM$  is not  $p$ -sensitive  $k$ -anonymous, and there is a masked microdata  $MM$  that satisfies constrained  $p$ -sensitive  $k$ -anonymity. This means that every  $QI$ -cluster from  $MM$  has at least  $k$  elements, it has at least  $p$  distinct values for every quasi-identifier attribute, and it does not have any constraint violation. Let  $cl_i$  be a cluster of elements from  $IM$  that is generalized to a  $QI$ -cluster in  $MM$  ( $i = 1, \dots, v$ ). Because  $MM$  satisfies constrained  $p$ -sensitive  $k$ -anonymity, the generalization of tuples in  $cl_i$  to  $gen(cl_i)$  does not create any constraint violation. Based on Property 1, for each quasi-identifier attribute, all entities from  $cl_i$  share the same  $MAGVals$ . As a consequence, by generalizing all quasi-identifier attributes values to their corresponding  $MAGVals$  (this is the procedure to create the  $MAM$  microdata) all entities from the cluster  $cl_i$  (for all  $i = 1, \dots, v$ ) will be contained within the same  $QI$ -cluster of  $MAM$ . This means that each  $QI$ -cluster in  $MAM$  contains one or more  $QI$ -clusters from  $MM$  and its size will, then, be at least  $k$  and the number of distinct values for every quasi-identifier attribute will be at least  $p$ . In conclusion,  $MAM$  is  $p$ -sensitive  $k$ -anonymous, which is a contradiction with our initial assumption. // q.e.d.

**Property 3.** If  $MAM$  satisfies  $p$ -sensitive  $k$ -anonymity then  $MAM$  satisfies the constrained  $p$ -sensitive  $k$ -anonymity property.

**Proof.** This follows from the definition of  $MAM$ .

**Property 4.** An initial microdata,  $IM$ , can be masked to comply with constrained  $p$ -sensitive  $k$ -anonymity using only generalization if and only if its corresponding  $MAM$  satisfies  $p$ -sensitive  $k$ -anonymity.

**Proof.**

“If”: If  $MAM$  satisfies  $p$ -sensitive  $k$ -anonymity, then based on Property 3,  $MAM$  is also constrained  $p$ -sensitive  $k$ -anonymous, and  $IM$  can be masked to  $MAM$  (in the worst case – or even to a less generalized masked microdata) to comply with constrained  $p$ -sensitive  $k$ -anonymity.

“Only If”: If  $MAM$  does not satisfy  $p$ -sensitive  $k$ -anonymity, then based on Property 2, any masked microdata obtained by applying generalization only to  $IM$  will not satisfy constrained  $p$ -sensitive  $k$ -anonymity. // q.e.d.

It is very likely that there are some  $QI$ -clusters in  $\mathcal{MAM}$  with size less than  $k$  or with less than  $p$  distinct values for a sensitive attribute. The entities belonging to these clusters cannot be masked to  $p$ -sensitive  $k$ -anonymity while preserving the constraint conditions, as shown by Property 6. We will use the notation  $OUT$  to represent these entities (for simplicity we use the same notation to refer to entities from both  $IM$  and  $\mathcal{MAM}$ ). For a given  $IM$  with its corresponding  $\mathcal{MAM}$  and  $OUT$  sets the following two properties hold.

**Property 5.**  $IM - OUT$  can be masked using generalization only to comply with constrained  $p$ -sensitive  $k$ -anonymity.

**Proof.** By definition of the  $OUT$  set, all  $QI$ -clusters from  $\mathcal{MAM} \setminus OUT$  have size  $k$  or more and any such cluster will have  $p$  distinct values for each quasi-identifier attribute, which means that  $\mathcal{MAM} \setminus OUT$  satisfies the  $p$ -sensitive  $k$ -anonymity property. Based on Property 4 ( $\mathcal{MAM} \setminus OUT$  is the maximum allowed microdata for  $IM \setminus OUT$ ),  $IM \setminus OUT$  can be masked using generalization only to comply with constrained  $p$ -sensitive  $k$ -anonymity. // q.e.d.

**Property 6.** Any subset of  $IM$  that contains one or more entities from  $OUT$  cannot be masked using generalization only to achieve constrained  $p$ -sensitive  $k$ -anonymity.

**Proof.** We assume that there is an initial microdata  $IM'$ , a subset of  $IM$ , that contains one or more entities from  $OUT$ , and  $IM'$  can be masked using generalization only to comply with constrained  $p$ -sensitive  $k$ -anonymity. Let  $x \in OUT \cap IM'$ . Let  $\mathcal{MAM}'$  be the maximum allowed microdata for  $IM'$ . Based on Property 4, if  $IM'$  can be masked to be constrained  $p$ -sensitive  $k$ -anonymous, then  $\mathcal{MAM}'$  is  $p$ -sensitive  $k$ -anonymous, therefore  $x$  will belong to a  $QI$ -cluster with size at least  $k$  and with  $p$  distinct values for every quasi-identifier attribute. By construction  $\mathcal{MAM}'$  is a subset of  $\mathcal{MAM}$ , and therefore, the size of each  $QI$ -cluster from  $\mathcal{MAM}$  is equal to or greater than the size of the corresponding  $QI$ -cluster from  $\mathcal{MAM}'$ . This means that  $x$  will belong to a  $QI$ -cluster with size at least  $k$  and with  $p$  distinct values for every quasi-identifier attribute in the  $\mathcal{MAM}$ . This is a contradiction with  $x \in OUT$ . // q.e.d.

Properties 5 and 6 show that  $OUT$  is the minimal tuple set that must be suppressed from  $IM$  such that the remaining set could be constrained  $p$ -sensitive  $k$ -anonymized. To compute a constrained  $p$ -sensitive  $k$ -anonymous masked microdata we apply the following steps. First, we suppress all tuples from the  $OUT$  set. Next, we create all  $QI$ -clusters in the maximum allowed microdata for  $IM - OUT$ . Last, each such cluster will be divided further, if possible, using the clustering approach from [1], [3], and [4], into several clusters, all with size

greater than or equal to  $k$  and with  $p$  or more distinct values for every sensitive attribute. This approach uses a greedy technique that tries to optimize an information loss ( $IL$ ) measure and a diversity measure.

At this moment there are several information loss measures in the literature such as the discernability metric [2], the normalized average cluster size metric [15], the utility assessment metrics [8], the information-theoretic measure of utility [11], and several variants of the information loss metric [3, 8, 29]. Any of the above measures could be used in our algorithm.

The information loss measure we choose to use in our algorithm implementation was introduced in [3]. We present it in Definition 10. We extend this measure, first, by computing the total information loss for a partition into clusters of the initial microdata set and, next, by normalizing it to the interval  $[0, 1]$  (see Definition 11). Note that this  $IL$  definition assumes that value generalization hierarchies are predefined for all quasi-identifier attributes.

**Definition 10 (Cluster Information Loss).** Let  $cl$  be a cluster of tuples from  $IM$ ,  $gen(cl)$  its generalization information and  $QI = \{Q_1, Q_2, \dots, Q_t\}$  the set of quasi-identifier attributes. The cluster information loss caused by generalizing  $cl$  tuples to  $gen(cl)$  is:

$$IL(cl) = |cl| \cdot \sum_{j=1}^t \frac{height\left(\Lambda(gen(cl)[Q_j])\right)}{height(H_{Q_j})},$$

where:

- $|cl|$  denotes the  $cl$  cluster's cardinality;
- $\Lambda(\mu)$ ,  $\mu \in \mathcal{H}_{Q_j}$  is the subhierarchy of  $\mathcal{H}_{Q_j}$  rooted in  $\mu$ ;
- $height(\mathcal{H}_{Q_j})$  denotes the height of the tree hierarchy  $\mathcal{H}_{Q_j}$ ;
- $t$  is the number of quasi-identifier attributes.

**Definition 11 (Normalized Total Information Loss).** The *normalized total information loss* for a partition into clusters,  $S$ , of the initial microdata set is the sum of the information loss for all clusters in  $S$  divided to the number of tuples from  $IM$  times the number of quasi-identifier attributes. Formally:

$$NTIL(IM, S) = \frac{\sum_{j=1}^{p+1} IL(cl_j)}{n \cdot t},$$

where:

- $n$  is the number of tuples from  $IM$ ;
- $t$  is the number of quasi-identifier attributes;

- $v + 1$  is the number of clusters from  $\mathcal{S}$ . The solution  $\mathcal{S}$  contain  $v$  clusters that are kept in the  $\mathcal{MM}$  and one cluster ( $cl_{v+1}$ ) that contains all entities from  $O\mathcal{UT}$  which are suppressed from the  $\mathcal{MM}$ .

It is worth noting that the cluster of tuples to be suppressed,  $cl_{v+1}$ , will have the maximum possible  $IL$  value for a cluster of the same size as  $cl_{v+1}$ . The information loss for this cluster will be:  $IL(cl_{v+1}) = |cl_{v+1}| \cdot t$ . When performing experiments (see Section 4) the information loss of the constrained anonymization solutions includes the information loss caused by the suppressed cluster as well, and not only for the generalized clusters. This way, the quality of the constrained  $p$ -sensitive  $k$ -anonymous solutions will not be biased because of a favored way of computing information loss for the suppressed tuples.

The cluster diversity measure, presented next, quantifies the heterogeneity degree of a cluster's tuples with respect to the sensitive attributes [4].

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

**Definition 12 (Diversity of Two Tuples).** The *diversity of two tuples*,  $X_i$  and  $X_j$  w.r.t. the sensitive attributes is given by:

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

where:

- $\delta(s_l^i, s_l^j) = \begin{cases} 1, & \text{if } s_l^i \neq s_l^j \\ 0, & \text{if } s_l^i = s_l^j \end{cases}$ ;
- $r$  is the number of sensitive attributes;
- $\sum_{l=1}^r w_l = 1$  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  $\mathcal{IM}$ .

**Definition 13 (Diversity between a tuple and a cluster).** The *diversity between a tuple  $X_i$  and a cluster  $cl$*  is given by

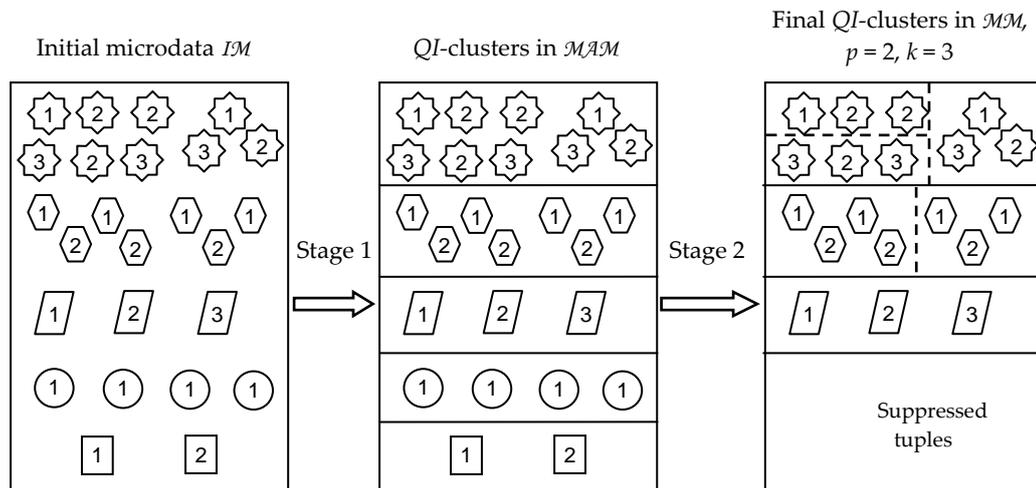
$$diversity(X_i, cl) = \sum_{l=1}^r w_l \cdot \delta(s_l^i, cl_l),$$

where:

- $\delta(s_l^i, cl_l) = \begin{cases} 1, & \text{if } s_l^i \text{ does not exist between the } S_l \text{ values in } cl \\ 0, & \text{if } s_l^i \text{ exists between the } S_l \text{ values in } cl \end{cases}$  ;
- $r$  is the number of sensitive attributes;
- the weights  $w_l$ ,  $\sum_{l=1}^r w_l = 1$ , have the same meaning as in Definition 12.

The two stage constrained  $p$ -sensitive  $k$ -anonymization algorithm called *GreedyCPKA* is depicted in Figure 3. The numbers correspond to different values of a sensitive attribute – we only consider one to make the process explainable graphically. The different geometrical shapes indicate tuples that belong to the same  $\mathcal{MAM}$  cluster. In the first step, the  $\mathcal{MAM}$  is formed. In the second step, a  $p$ -sensitive  $k$ -anonymization algorithm is applied on every  $\mathcal{MAM}$  cluster with more than  $k$  entities and  $p$  distinct values for each sensitive attribute. In this last step, the  $\mathcal{MAM}$  clusters that have less than  $k$  entities or less than  $p$  distinct values for each sensitive attribute are suppressed. The suppressed entities form the set  $OUT$ .

**Figure 3. The two-stage process in creating constrained  $p$ -sensitive  $k$ -anonymous masked microdata.**



---

We present next the pseudocode of the *GreedyCPKA* Algorithm.

**Algorithm GreedyCPKA is**

**Input**  $IM$  - initial microdata set;  
 $p, k$  - as in  $p$ -sensitive  $k$ -anonymity;  
generalization boundaries;

**Output**  $S = \{cl_1, cl_2, \dots, cl_v, cl_{v+1}\}$ ;  
 $\bigcup_{j=1}^{v+1} cl_j = IM$ ;  
 $cl_i \cap cl_j = \emptyset, i, j = 1..v+1, i \neq j; |cl_j| \geq k, cl_j$  is  $p$ -sensitive for every sensitive attribute, generalizing  $cl_j$  doesn't produce constraint violations,  $j=1..v$  (i.e.  $S - \{cl_{v+1}\}$  is a set of clusters that ensure constrained  $p$ -sensitive  $k$ -anonymity,  $cl_{v+1}$  is suppressed);

Compute  $MAM$  and  $OVT$ ;  
 $S = \emptyset$ ;  
For each  $cl \in MAM - OVT$  do  
// By  $cl$  we refer to the entities from  $IM$   
// that are clustered together in  $MAM$ .  
 $S' = \text{GreedyPKClustering}(cl, p, k)$ ;  
 $S = S \cup S'$ ;  
End For;  
 $v = |S|$ ;  
 $cl_{v+1} = OVT$ ;  
End **GreedyCPKA**.

**Function GreedyPKClustering( $IM, p, k$ )**

$S = \emptyset; i = 1$ ;  
 $X_{seed}$  = a randomly selected tuple from  $IM$ ;  
Repeat  
 $cl_i = \emptyset$ ;  
 $X_{seed} = \text{one tuple from } \underset{X \in IM}{\text{argmax diversity}}(X_{seed}, X)$  ;  
// one of the most diverse tuples  $\in IM$  w.r.t. old  $X_{seed}$   
 $cl_i = cl_i \cup \{X_{seed}\}$ ;  
 $IM = IM - \{X_{seed}\}$ ;  
Repeat  
 $Temp = \underset{X \in IM}{\text{argmax diversity}}(X, cl_i)$ ;  
 $X^* = \text{one tuple from } \underset{X \in Temp}{\text{argmin IL}}(X, cl_i)$ ;  
// one tuple within the most diverse tuples w.r.t.  $cl_i$   
// that produce the minimal  $IL$  growth when added to  $cl_i$   
 $cl_i = cl_i \cup \{X^*\}$ ;  
 $IM = IM - \{X^*\}$ ;  
Until ( $cl_i$  is  $p$ -sensitive) or ( $IM = \emptyset$ );

---

```

If ( $|cl_i| < k$ ) and ( $IM \neq \emptyset$ ) then
  Repeat
     $X^* = \text{one tuple from } \underset{X \in IM}{\text{argmin}} IL(cl_i \cup \{X\})$ 
     $cl_i = cl_i \cup \{X^*\};$ 
     $IM = IM - \{X^*\};$ 
  Until ( $cl_i$  is  $k$ -anonymous) or ( $IM = \emptyset$ );
End If;
If ( $|cl_i| \geq k$  and  $cl_i$  is  $p$ -sensitive) then
   $S = S \cup \{cl_i\};$ 
   $i++;$ 
Else
   $\text{DisperseCluster}(S, cl_i);$ 
  // only last cluster can be dispersed
End If;
Until  $IM = \emptyset$ ;
Return  $S$ ;
End GreedyPKClustering.

```

```

Function DisperseCluster( $S, cl$ )
  For each  $X \in cl$  do
     $cl^* = \text{FindBestCluster}(X, S);$ 
     $cl^* = cl^* \cup \{X\};$ 
  End For;
End DisperseCluster.

```

```

Function FindBestCluster( $X, S$ ) is
  bestCluster = null;
  infoLoss =  $\infty$ ;
  For each  $cl_j \in S$  do
    If  $IL(cl_j \cup \{X\}) < \text{infoLoss}$  then
      infoLoss =  $IL(cl_j \cup \{X\})$ ;
      bestCluster =  $cl_j$ ;
    End If;
  End For;
  Return bestCluster;
End FindBestCluster.

```

## 4 Experimental Results

In this section we compare the *GreedyCPKA* (generates constrained  $p$ -sensitive  $k$ -anonymous masked microdata), *GreedyCKA* (generates constrained  $k$ -anonymous masked microdata) [23] and *GreedyPKClustering* (generates unconstrained  $p$ -sensitive  $k$ -anonymous masked microdata) [4] algorithms, from different perspectives:

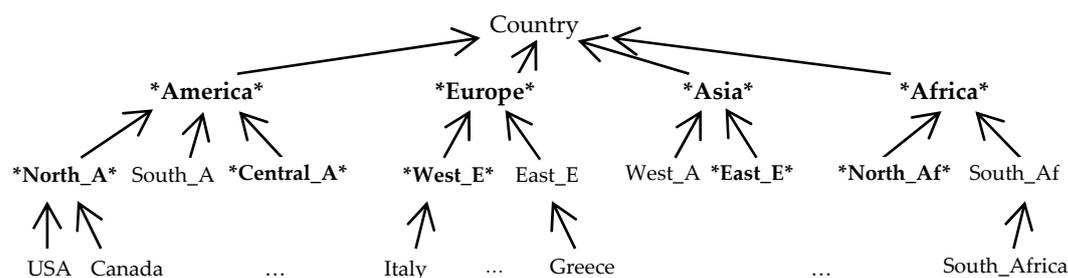
- The quality of the results they produce measured according to the normalized total information loss metric;
- The algorithms' efficiency as expressed by their running time;
- The number of constraint violations that  $p$ -sensitive  $k$ -anonymous masked microdata produced by *GreedyPKClustering* have; and,
- The suppression amount performed by *GreedyCPKA* in order to produce constrained  $p$ -sensitive  $k$ -anonymous microdata in the presence of constraints.

The algorithms were implemented in Java, and tests were executed on a dual CPU machine with 3.0 GHz and 4 GB of RAM.

The experiments were performed on the *Adult* dataset from the UC Irvine Machine Learning Repository [25] consisting of 45222 tuples. In all the experiments we considered a set of six quasi-identifier attributes: *age*, *workclass*, *marital\_status*, *race*, *sex*, and *native\_country*, and a set of three sensitive attributes: *education\_num*, *education*, and *occupation*. The three algorithms were applied to this microdata set for  $k = 4, 8, 10$  and  $20$  and  $p = 2, 3, 4, 6, 8, 10$ , and  $13$  (only when  $p$  is less or equal to  $k$ ). The sensitive attribute weights were set in all experiments to  $0.3$  (for *education* and *education\_num*) and  $0.4$  (for *occupation*). The weights are used for assessing the diversity with respect to sensitive attributes between tuples and clusters (see Definitions 12 and 13).

We considered generalization boundaries for the *age* and *native\_country* quasi-identifiers. The value hierarchy and the *MAGVals* for *age* are as presented by Figure 2c); *native\_country* has a 4-level constrained hierarchy (see Figure 4). The quasi-identifier attributes without constraint boundaries have the following heights for their generalization hierarchies: *workclass*-1, *sex*-1, *race*-1, and *marital\_status*-2.

Figure 4. *MAGVals* for the quasi-identifier attribute *native\_country*.



The experiments we performed show that information loss is higher when  $p$ -sensitive  $k$ -anonymity (constrained or not) is enforced on a dataset compared to when the dataset is masked according to  $k$ -anonymity only (Figure 5). Of course, this is a result to be expected, as there is always a tradeoff between pre-

servicing data utility and protecting it with a more powerful anonymity model. On the other hand, Figure 5 also proves that the information loss does not degrade significantly when constraints are incorporated into  $p$ -sensitive  $k$ -anonymity: *GreedyCPKA* and *GreedyPKClustering* perform similarly with respect to information loss.

Figure 5. Normalized Total Information Loss (NTIL) for *GreedyCPKA*, *GreedyPKClustering*, and *GreedyCKA*.

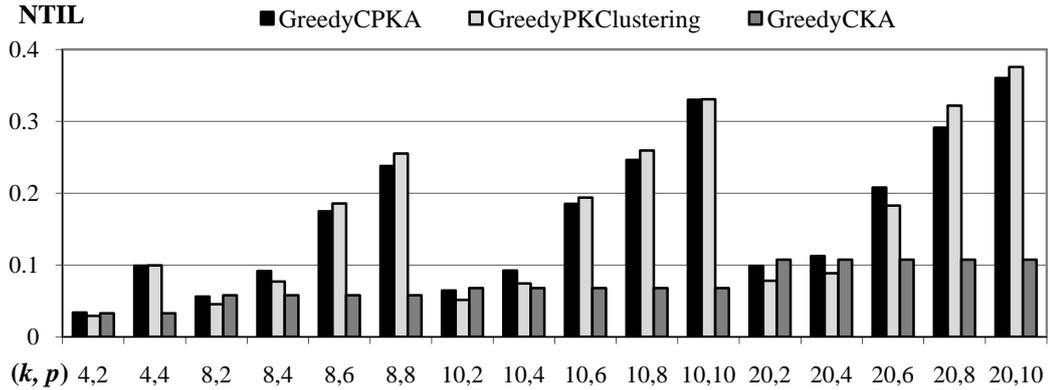
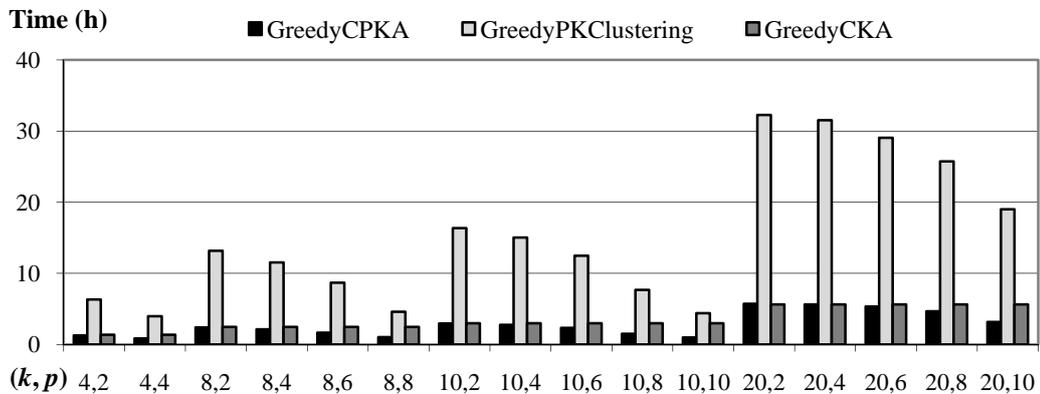


Figure 6. Running Time for *GreedyCPKA*, *GreedyPKClustering*, and *GreedyCKA*.



Our results show that *GreedyCPKA* and *GreedyCKA* perform equally efficiently on all  $p$  and  $k$  values, including high values, where *GreedyPKClustering* slows down significantly (Figure 6).

When the unconstrained *GreedyPKClustering* is applied to anonymize  $IM$ , the resulting masked microdata usually contains numerous constraint violations. Table 5 reports the number of constraint violations in some outcomes of *GreedyPKClustering*. Also, *GreedyCKA* alone will obviously produce constrained  $k$ -anonymous masked microdata that is very likely to not meet the  $p$ -sensitivity criterion and, therefore, it will allow attribute disclosures.

---

$k$	$p$	# constraint violations
4	2	3958
4	3	6286
4	4	12743
8	2	7260
8	3	7766
8	4	10836
8	6	21213
8	8	27017
10	2	8686
10	3	9164
10	4	11345
10	6	22853
10	8	27882
10	10	36073
20	2	13990
20	3	14171
20	4	14896
20	6	26893
20	8	36437
20	10	43058
20	13	61391

---

**Table 5. Constraint violations in *GreedyPKClustering* outcomes**

Our experiments also show that the number of suppressed tuples increases as  $k$  and  $p$  values increase. This is a result of having  $\mathcal{MAM}$ -clusters with few tuples (not achieving  $k$ ) or low diversity (not achieving  $p$ ). Table 6 summarizes the number of suppressed tuples for both *GreedyCPKA* and *GreedyCKA* algorithms. It is worth noting that this number is low compared with total number of tuples (45,222).

---

$k$	$p$	# suppressed tuples ( <i>GreedyCPKA</i> )	# suppressed tuples ( <i>GreedyCKA</i> )
4	2	11	11
4	3	17	11
4	4	38	11
8	2	42	42
8	3	42	42
8	4	52	42
8	6	127	42
8	8	358	42
10	2	51	51
10	3	51	51
10	4	61	51
10	6	136	51
10	8	358	51
10	10	463	51
20	2	162	162
20	3	162	162
20	4	162	162
20	6	192	162
20	8	370	162
20	10	463	162
20	13	1593	162

---

Table 6. Suppressed tuples in *GreedyCPKA* and *GreedyCKA* outcomes

## 5 Related Work

The anonymization problem for microdata was widely studied in the research literature [1, 2, 3, 14, 15, 26, 27, 28]. Many anonymization models that protect confidential information were proposed [7, 16, 21, 22, 29, 31, 32, 33, 35, 26]. Depending on specific model, heuristic or optimal anonymization algorithms were introduced. So far, the research community focused on minimizing a measure of information loss while guaranteeing the privacy requirements. This problem is called *the privacy-constrained anonymization problem* [10]. This may lead to distorting the data such that it becomes unusable to research users. To prevent this situation a new approach was long due, the utility/information loss

---

constraints must be specified prior to the anonymization process and the privacy requirements (usually an anonymity model) must be enforced within the given constraints.

The first paper that considered this approach was [23]. In this paper the *constrained k-anonymity*, a privacy model that preserves the  $k$ -anonymity requirement while specifying quasi-identifiers generalization boundaries (or limits), was introduced. The downfall of this model is that it does not protect against attribute disclosure [13]. The current paper addresses this issue by incorporating both generalization constraints and a privacy model that protects against attribute disclosure.

Concurrently, Loukides, Tziatzios, and Shao introduced the *preference-constrained k-anonymization* [19]. They allow data owners to specify usage requirements as a set of preferences on attributes or data values (in other words forming a set of constraints) and solve the anonymization problem as a multi-objective optimization problem. There are two types of constraints defined in this work: the *attribute level preference* and the *value-level preference*. The attribute-level preference is similar to defining all *MAGVals* at the same level in the value generalization hierarchy tree; the value-level preference resembles the constraints definition from this paper. This paper also defines preferences for numerical attributes in terms of ranges. An important difference between [19] and the current paper is again the selection of the privacy model ( $k$ -anonymity versus a model that protects against both identity and attribute disclosure).

Constraints that limit the amount of distortion for transaction anonymization are introduced by Loukides, Gkoulalas-Divanis, and Malin [20]. In this paper, the authors focus on both privacy and utility constraints in the context of transactions. The authors argue the importance of including utility constraints from the early stages in the anonymization process. This paper focuses on a different data model (transactional data) and only on  $k$ -anonymity.

Recently, Ghinita, Karras, and Kalinis introduced the *accuracy-constrained anonymization* problem [10]. This problem finds the maximum degree of privacy (for either  $k$ -anonymity or  $l$ -diversity) that can be achieved such that the information loss (the definition of  $IL$  can vary) does not exceed a threshold  $E$ . This work is related to the one presented in this paper since it is the only other work that discusses the utility constraints in the context of an anonymization model that protects against attribute disclosure. The main difference between their work and ours resides in the way in which constraints are enforced. We see these constraints as dependent on individual values, and this is why we associate *MAGVals* to each possible value for a constrained quasi-identifier attribute. In [10], the constraints are global in terms of a general information loss measure. This limits the flexibility in defining useful value-based constraints.

In a preliminary two-page version of this paper, Campan, Truta, and Cooper [5] first introduced constrained  $p$ -sensitive  $k$ -anonymity. Compared to its preliminary version, the current paper presents the complete theoretical framework for this new model, the description of the *GreedyCPKA* Algorithm, and an array of experimental results.

The research papers that include utility constraints in the process of anonymization are in general complementary; there are differences in how these constraints are defined and what anonymization model is used. Our approach incorporates a very practical set of constraints (generalization boundaries) with a privacy model that protects against both identity disclosure and attribute disclosure.

## 6 Conclusions

In this paper we presented a new privacy model that protects against both identity and attribute disclosure while keeping the quasi-identifiers generalization restricted to certain user-specified boundaries. We also introduced a greedy algorithm that will generate a masked microdata which will conform to the new privacy model. The experiments show that the proposed algorithm is comparable with existing algorithms used for generating  $p$ -sensitive  $k$ -anonymity with respect to the results' quality, and the obtained masked microdata complies with the generalization boundaries as indicated by the user. A last remark needs to be made: the concept of constrained anonymity can certainly be translated from  $p$ -sensitive  $k$ -anonymity to other models such as  $l$ -diversity,  $t$ -closeness, etc. that are derived from  $k$ -anonymity.

## Acknowledgements

This work was partially supported by a grant from the Center from Integrative Science and Mathematics (CINSAM) from Northern Kentucky University.

---

## References

- [1] Aggarwal, G., Feder, T., Kenthapadi, K., Motwani, R., Panigrahy, R., Thomas, D., and Zhu, A. (2006) Achieving Anonymity via Clustering. In *Proceedings of the ACM PODS 2006 Conference*, pages 153 – 162.
- [2] Bayardo, R.J. and Agrawal, R. (2005) Data Privacy through Optimal  $k$ -Anonymization. In *Proceedings of the IEEE International Conference of Data Engineering*, pages 217 – 228.
- [3] Byun, J.W., Kamra, A., Bertino, E., and Li, N. (2007) Efficient  $k$ -Anonymization using Clustering Techniques. In *Proceedings of the 11th International Conference on Database Systems for Advanced Applications (DAS-FAA 2007)*, pages 188 – 200.
- [4] Campan, A., Truta, T. M., Miller, J., and Sinca, R. (2007) A Clustering Approach for Achieving Data Privacy. In *Proceedings of the 2007 International Conference on Data Mining (DMIN'07)*, pages 321 – 327.
- [5] Campan, A., Truta, T.M., and Cooper, N. (2010) User-Controlled Generalization Boundaries for  $P$ -Sensitive  $K$ -Anonymity. In *Proceedings of the ACM Symposium on Applied Computing (SAC2010)*.
- [6] Chen, B.C., Ramakrishnan, R., and LeFevre, K. (2007) Privacy Skyline: Privacy with Multidimensional Adversarial Knowledge. In *Proceedings of the Very Large Data Bases*, pages 770 – 781.
- [7] Domingo-Ferrer, J., Sebe, F., and Solanas, A. (2008) An Anonymity Model Achievable via Microaggregation. In *Proceedings of the 5th VLDB Workshop on Secure Data Management*, pages 209 – 218.
- [8] Fouad, M.R., Lebanon, G., and Bertino, E. (2008) ARUBA: A Risk-Utility-Based Algorithm for Data Disclosure. In *Proceedings of the 5th VLDB Workshop on Secure Data Management*, pages 32 – 49.
- [9] Ghinita, G., Karras, P., Kalinis, P., and Mamoulis, N. (2007) Fast Data Anonymization with Low Information Loss. In *Proceedings of the Very Large Data Base Conference*, pages 758 – 769.
- [10] Ghinita, G., Karras, P., Kalinis, P., and Mamoulis, N. (2009) A Framework for Efficient Data Anonymization under Privacy and Accuracy Constraints. *ACM Transactions on Database Systems (TODS)*, Article number 9.
- [11] Goldberger, J. and Tassa, T. (2009) Efficient Anonymizations with Enhanced Utility. In *Proceedings of the International Workshop on Privacy Aspects of Data Mining*.
- [12] HIPAA (2002) Health Insurance Portability and Accountability Act. Online at <http://www.hhs.gov/ocr/hipaa>.

- 
- [13] Lambert, D. (1993) Measures of Disclosure Risk and Harm. *Journal of Official Statistics*, Vol. 9, pages 313 – 331.
- [14] LeFevre, K., DeWitt, D., and Ramakrishnan, R. (2005) Incognito: Efficient Full-Domain  $K$ -Anonymity. In *Proceedings of the ACM SIGMOD 2005 Conference*, pages 49 – 60.
- [15] LeFevre, K., DeWitt, D., and Ramakrishnan, R. (2006) Mondrian Multidimensional  $K$ -Anonymity. In *Proceedings of the IEEE International Conference of Data Engineering*, pages 25.
- [16] Li, N., Li, T., and Venkatasubramanian, S. (2007)  $T$ -Closeness: Privacy Beyond  $k$ -Anonymity and  $l$ -Diversity. In *Proceedings of the 23rd International Conference on Data Engineering (IEEE ICDE 2007)*, pages 106 – 115.
- [17] Li, J, Tao, Y., and Xiao, X. (2008) Preservation of Proximity Privacy in Publishing Numerical Sensitive Data. In *Proceedings of the ACM SIGMOD*, pages 473 – 486.
- [18] Liu, J. Q. and Wang, K. (2010) On Optimal Anonymization for  $l^+$ -Diversity. In *Proceedings of the International Conference on Data Engineering (IEEE ICDE 2010)*.
- [19] Loukides, G., Tziatzios, A., and Shao, J. (2009) Towards Preference-Constrained  $k$ -Anonymisation. In *DASFAA Workshops 2009*, pages 231 – 245.
- [20] Loukides, G., Gkoulalas-Divanis, A., and Malin, B. (2009) CONstraint-based Anonymization of Transactions. In *CoRR abs/0912.2548*.
- [21] Machanavajjhala, A., Gehrke, J., and Kifer, D. (2006)  $L$ -Diversity: Privacy beyond  $K$ -Anonymity. In *Proceedings of the International Conference on Data Engineering (IEEE ICDE 2006)*, pages 24.
- [22] Martin, D.J., Kifer, D., Machanavajjhala, A., and Gehrke, J. (2007) Worst-Case Background Knowledge for Privacy-Preserving Data Publishing. In *Proceedings of the IEEE International Conference on Data Engineering*, pages 126 – 135.
- [23] Miller, J., Campan, A., and Truta, T. M. (2008) Constrained  $K$ -Anonymity: Privacy with Generalization Boundaries. In *Proceedings of the Workshop on Practical Preserving Data Mining, with SIAM SDM 2008*.
- [24] Nergiz, M.E., Atzori, M., and Clifton, C. (2007) Hiding the Presence of Individuals from Shared Databases. In *Proceedings of the ACM SIGMOD*, pages 665 – 697.
- [25] Newman, D. J., Hettich, S., Blake, C. L., and Merz, C. J. (1998) UCI Repository of Machine Learning Databases. Online at [www.ics.uci.edu/~mllearn/MLRepository.html](http://www.ics.uci.edu/~mllearn/MLRepository.html), UC Irvine.
-

- 
- [26] Samarati, P. (2001) Protecting Respondents Identities in Microdata Release. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 13, No. 6, pages 1010 – 1027.
- [27] Sweeney, L. (2002)  $k$ -Anonymity: A Model for Protecting Privacy. *International Journal on Uncertainty, Fuzziness, and Knowledge-based Systems*, Vol. 10, No. 5, pages 557 – 570.
- [28] Sweeney, L. (2002) Achieving  $k$ -Anonymity Privacy Protection Using Generalization and Suppression. *International Journal on Uncertainty, Fuzziness, and Knowledge-based Systems*, Vol. 10, No. 5, pages 571 – 588.
- [29] Truta, T.M. and Bindu, V. (2006) Privacy Protection:  $P$ -Sensitive  $K$ -Anonymity Property. In *Proceedings of the Workshop on Privacy Data Management, with ICDE 2006*, pages 94.
- [30] Truta, T.M. and Campan, A. (2007)  $K$ -Anonymization Incremental Maintenance and Optimization Techniques. In *Proceedings of the ACM Symposium of Applied Computing*, 380 – 387.
- [31] Xiao, X. and Tao, Y. (2006) Personalized Privacy Preservation. In *Proceedings of the ACM SIGMOD*, pages 229 – 240.
- [32] Xiao, X. and Tao, Y. (2007)  $m$ -Invariance: Towards Privacy Preserving Republication of Dynamic Datasets. In *Proceedings of the ACM SIGMOD*, pages 689 – 700.
- [33] Wei, Q., Lu, Y. and Lou, Q. (2008)  $(\tau, \lambda)$ -Uniqueness: Anonymity Management for Data Publication. In *Proceedings of the IEEE International Conference on Computer and Information Science*.
- [34] Winkler, W. (1995) Matching and Record Linkage. In *Business Survey Methods*, Wiley, pages 374 – 403.
- [35] Wong, R.C.W., Li, J., Fu, A.W.C., and Wang, K. (2006)  $(\alpha, k)$ -Anonymity: An Enhanced  $k$ -Anonymity Model for Privacy-Preserving Data Publishing. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD 2006)*, pages 754 – 759.
- [36] Wong, R.C.W., Li, J., Fu, A.W.C. and Pei, J. (2007) Minimality Attack in Privacy-Preserving Data Publishing. In *Proceedings of the Very Large Data Bases*, pages 543 – 554.
- [37] Zhang, Q., Koudas, N., Srivastava, D., and Yu, T. (2007) Aggregate Query Answering on Anonymized Tables. In *Proceedings of the IEEE International Conference on Data Engineering (IEEE ICDE 2007)*, pages 116 – 125.