Constrained *k*-Anonymity: Privacy with Generalization Boundaries

John Miller[*] Alina Campan[*, $] Traian Marius Truta[*]

Abstract: In the last few years, due to new privacy regulations, research in data privacy has flourished. A large number of privacy models were developed most of which are based on the *k*-anonymity property. Because of several shortcomings of the *k*-anonymity model, other privacy models were introduced (*l*-diversity, *p*-sensitive *k*-anonymity, (α, *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 to the initial data are imposed in order to preserve its usefulness. Our approach consists of specifying quasi-identifiers generalization boundaries, and achieving *k*-anonymity within the imposed boundaries. We think that limiting the amount of generalization when masking microdata is indispensable for real life datasets and applications. In this paper, the *constrained k-anonymity* model and its properties are introduced and an algorithm for generating constrained *k*-anonymous microdata is presented. Our experiments have shown that the proposed algorithm is comparable with existing algorithms used for generating *k*-anonymity with respect to results quality, and that by using existing unconstrained *k*-anonymization algorithms the generalization boundaries are violated. We also discuss how the constrained *k*-anonymity model can be easily extended to other privacy models.

1 Introduction

A huge interest in data privacy has been generated recently within the public and media [14], as well as in the legislative body [6] and research community.

Many research efforts have been directed towards finding methods to anonymize datasets to satisfy the *k*-anonymity property [16, 17]. These methods also consider minimizing one or more cost metrics between the initial and released microdata (a dataset where each tuple corresponds to one individual entity). Of particular interest are the cost metrics that quantify the *information loss* [2, 5, 19, 27]. Although producing the optimal solution for the *k*-anonymity problem w.r.t. various proposed cost measures has been proved to be NP-hard [9], there are several polynomial algorithms that produce good solutions for the *k*-anonymity problem for real life datasets [1, 2, 8, 9, 21].

Recent results have shown that *k*-anonymity fails to protect the privacy of individuals in all situations [12, 20, 26]. Several privacy models that extend the *k*-anonymity model have been proposed in the literature to avoid *k*-anonymity short-comings: *p*-sensitive *k*-anonymity [20] with its extension called extended *p*-sensitive *k*-anonymity [3], *l*-diversity [12], (α, *k*)-anonymity [24], *t*-closeness [10], (*k*, *e*)-anonymity [28], (*c*, *k*)-safety [13], *m*-confidentiality [25], personalized privacy [26], etc.

In general, the existing anonymization algorithms use different quasi-identifiers generalization strategies in order to obtain a masked microdata that is *k*-anonymous (or satisfies an extension of *k*-anonymity) and conserves as much information intrinsic to the initial microdata as possible. To our knowledge, a privacy model that considers the specification of the maximum allowed generalization level for quasi-identifier attributes in the masked microdata does not exist, nor does a corresponding anonymization algorithm capable of controlling the generalization amount. The ability to limit the amount of allowed generalization could be valuable, and, in fact, indispensable for real life datasets. For example, for some specific data analysis tasks, available masked microdata with the address information generalized beyond the US state level could be useless. In this case the only solution would be to ask the owner of the initial microdata to have the anonymization algorithm applied repeatedly on that data, perhaps with a decreased level of anonymity (a smaller *k*) until the masked microdata satisfies the maximum generalization level requirement (i.e. no address is generalized further than the US state).

In this paper, we first introduce a new anonymity model, called ***constrained k-anonymity***, which preserves the *k*-anonymity requirement while specifying quasi-identifiers generalization boundaries

P3DM'08, April 26, 2008, Atlanta, Georgia, USA.

[*] Department of Computer Science, Northern Kentucky University, U.S.A., {millerj10, campana1, trutat1}@nku.edu

[$] Visiting from Department of Computer Science, Babes-Bolyai University, Romania

(or limits). Second, we describe an algorithm to transform a microdata set such that its corresponding masked microdata will comply with the constrained *k*-anonymity. This algorithm relies on several properties stated and proved for the proposed privacy model.

The paper is organized as follows. Section 2 introduces basic data privacy concepts; and generalization and tuple suppression techniques as a mean to achieve data privacy. Section 3 presents the new constrained *k*-anonymity model. An anonymization algorithm to transform microdata to comply with constrained *k*-anonymity is described in Section 4. Section 5 contains comparative quality results, in terms of information loss, processing time, for our algorithm and one of the existing *k*-anonymization algorithms. The paper ends with future work directions and conclusions.

2 *K*-Anonymity, Generalization and Suppression

Let IM be the initial microdata and MM be the released (a.k.a. masked) microdata. The attributes characterizing IM are classified into the following three categories:

- *identifier* attributes such as *Name* and *SSN* that can be used to identify a record.
- *key* or *quasi-identifier* attributes such as *ZipCode* and *Age* that may be known by an intruder.
- *sensitive* or *confidential* attributes such as *PrincipalDiagnosis* and *Income* that are assumed to be unknown to an intruder.

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 is that the values for the confidential attributes are not available from any external source. This assumption guarantees that an intruder cannot use the confidential attributes' values to increase his/her chances of disclosure, and, therefore, modifying this type of 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.

To rigorously and succinctly express the *k*-anonymity property, we use the following concept:

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

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* [26].

We define *k*-anonymity based on the minimum size of all *QI*-clusters.

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

A general method widely used for masking initial microdata to conform to the *k*-anonymity model is the generalization of the quasi-identifier attributes. Generalization consists in replacing the actual value of the attribute with a less specific, more general value that is faithful to the original [18].

Initially, this technique was used for *categorical* attributes and employed predefined domain and value generalization hierarchies [18]. Generalization was extended for *numerical* attributes either by using *pre-defined hierarchies* [7] or a *hierarchy-free model* [9].

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 *Sex*.

There are several ways to perform generalization. Generalization that maps all values of a quasi-identifier categorical attribute from IM to a more general domain in its domain generalization hierarchy is called *full-domain generalization* [9, 16]. Generalization can also map an attribute's values to different domains in its domain generalization hierarchy, each value being replaced by the same generalized value in the entire dataset [7]. The least restrictive generalization, called *cell level generalization* [11], extends Iyengar model [7] by allowing the same value to be mapped to different generalized values, in distinct tuples.

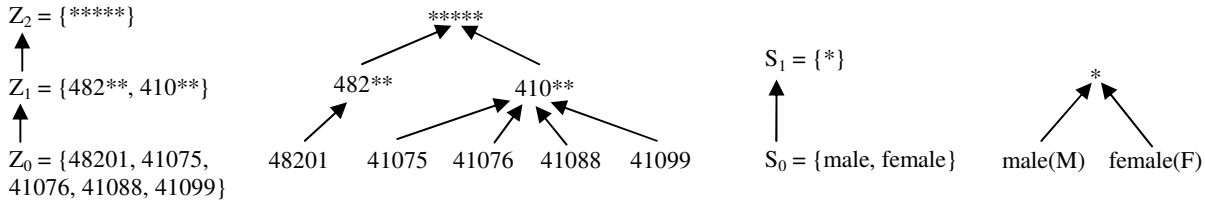

Figure 1: Examples of domain and value generalization hierarchies.

Tuple suppression [16, 18] is the only other method used in this paper for masking the initial microdata. By eliminating entire tuples we are able to reduce the amount of generalization required for achieving the *k*-anonymity property in the remaining tuples. Since the constrained *k*-anonymity model uses generalization boundaries, for many initial microdata sets suppression has to be used in order to generate constrained *k*-anonymous masked microdata.

3 Constrained *K*-Anonymity

In order to specify a generalization boundary, we introduce the concept of a maximum allowed generalization value that is associated with each possible quasi-identifier attribute value from IM. This concept is used to express how far the owner of the data thinks that the quasi-identifier's values could be generalized, such that the resulted masked microdata would still be useful. Limiting the amount of generalization for quasi-identifier attribute values is a necessity for various uses of the data. The data owner is often aware of the way various researchers are using the data and, as a consequence, he/she is able to identify maximum allowed generalization values. For instance, when the released microdata is used to compute various statistical measures related to the US states, the data owner will select the states as maximal allowed generalization values. The desired protection level should be achieved with minimal changes to the initial microdata IM. However, minimal changes may cause generalization that surpasses the maximal allowed generalization values and the masked microdata MM would become unusable. More changes are preferred in this situation if they do not contradict the generalization boundaries.

At this stage, for simplicity, we use predefined hierarchies for both categorical and numerical quasi-identifier attributes, when defining maximal allowed generalization values. Techniques to dynamically build hierarchies for numerical attributes exist in the literature [4] and we intend to use them in our future research.

Definition 3. (***Maximum Allowed Generalization Value***): Let Q be a quasi-identifier attribute (categorical or numerical), and $\mathcal{H}_Q$ its predefined value generalization hierarchy. For every leaf value $v \in \mathcal{H}_Q$, the ***maximum allowed generalization value*** of v, denoted by $MAGVal(v)$, is the value (leaf or not-leaf) in $\mathcal{H}_Q$ situated on the path from v to the root, such that:

- for any released microdata, the value v is permitted to be generalized only up to $MAGVal(v)$ and
- when several *MAGVal*s exist on the path between v and the hierarchy root, then the $MAGVal(v)$ is the first *MAGVal* that is reached when following the path from v to the root node.

Figure 2 contains an example of defining maximal allowed generalization values for a subset of values for the *Location* attribute. The *MAGVal*s for the leaf values "San Diego" and "Lincoln" are "California", and, respectively, "Midwest" (the *MAGVal*s are marked by * characters that delimit them). This means that the quasi-identifier *Location*'s value "San Diego" may be generalized to itself or "California", but not to "West Coast" or "United States". Also, "Lincoln" may be generalized to itself, "Nebraska", or "Midwest", but not to "United States".

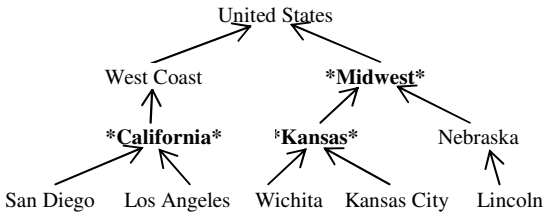

Figure 2: Examples of *MAGVal*s.

The second requirement in the *MAGVal*'s definition specifies that the hierarchy path between a leaf value v and $MAGVal(v)$ can contain no node other than $MAGVal(v)$ that is a maximum allowed generalization value. This restriction is imposed in order to avoid any ambiguity about the *MAGVal*s of the leaf values in a sensitive attribute hierarchy. Note that several *MAGVal*s may exist on a path between a leaf and the root as a result of defining *MAGVal*s for other leaves within that hierarchy.

Definition 4. (***Maximum Allowed Generalization Set***): Let Q be a quasi-identifier attribute and $\mathcal{H}_Q$ its predefined value generalization hierarchy. The set of all *MAGVal*s for attribute Q is called Q's ***maximum allowed generalization set***, and it is denoted by $MAGSet(Q) = \{ MAGVal(v) \mid \forall v \in leaves(\mathcal{H}_Q) \}$ (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 2, $MAGSet(Location) = \{$California, Kansas, 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.

Record	Name	SSN	Age	Location	Sex	Race	Diagnosis	Income
r_1	Alice	123456789	32	San Diego	M	W	AIDS	17,000
r_2	Bob	323232323	30	Los Angeles	M	W	Asthma	68,000
r_3	Charley	232345656	42	Wichita	M	W	Asthma	80,000
r_4	Dave	333333333	30	Kansas City	M	W	Asthma	55,000
r_5	Eva	666666666	35	Lincoln	F	W	Diabetes	23,000
r_6	John	214365879	20	Lincoln	M	B	Asthma	55,000
r_7	Casey	909090909	25	Wichita	F	B	Diabetes	23,000

Figure 3: An initial microdata set $\mathcal{IM}$

Record	Age	Location	Sex	Race
r_1	30-32	California	M	W
r_2	30-32	California	M	W
r_3	30-42	MidWest	*	W
r_4	30-42	MidWest	*	W
r_5	30-42	MidWest	*	W
r_6	20-25	MidWest	*	B
r_7	20-25	MidWest	*	B

a)

Record	Age	Location	Sex	Race
r_1	30-32	California	M	W
r_2	30-32	California	M	W
r_3	25-42	Kansas	*	*
r_4	25-42	Kansas	*	*
r_7	25-42	Kansas	*	*
r_5	20-35	Lincoln	*	*
r_6	20-35	Lincoln	*	*

b)

Figure 4: Two masked microdata sets $\mathcal{MM}_1$ and $\mathcal{MM}_2$ for the initial microdata $\mathcal{IM}$. (Only the quasi-identifier attribute values are shown in the masked microdata sets)

Definition 5. (***Constraint Violation***): We say that the masked microdata $\mathcal{MM}$ has a ***constraint violation*** if one quasi-identifier value, v, in $\mathcal{IM}$, is generalized in one tuple in $\mathcal{MM}$ beyond its specific maximal generalization value, $MAGVal(v)$.

Definition 6. (***Constrained K-Anonymity***): The masked microdata $\mathcal{MM}$ satisfies the ***constrained k-anonymity property*** if it satisfies k-anonymity and it does not have any constraint violation.

We note that a k-anonymous masked microdata may have multiple constraint violations, but any masked microdata that satisfies constrained k-anonymity property will not have any constraint violations; or in other words, any quasi-identifier value, v, from the initial microdata will never be generalized beyond its $MAGVal(v)$ in any constrained k-anonymous masked microdata.

Consider the following example. The initial microdata set $\mathcal{IM}$ in Figure 3 is characterized by the following attributes: *Name* and *SSN* are identifier attributes (to be removed from the masked microdata), *Age*, *Location*, *Sex*, and *Race* are the quasi-identifier attributes, and *Diagnosis* and *Income* are the sensitive attributes. The attribute *Location* values and their *MAGVal*s are described by Figure 2. The remaining quasi-identifier attributes do not have any generalization boundary requirements.

Figure 4 illustrates two possible masked microdata $\mathcal{MM}_1$ and $\mathcal{MM}_2$ for the initial microdata $\mathcal{IM}$. In this figure, only quasi-identifier values are shown, the confidential attribute values will be kept unchanged from the initial microdata $\mathcal{IM}$ (*Diagnosis* and *Income* attributes from Figure 3). The first masked microdata, $\mathcal{MM}_1$, satisfies 2-anonymity, but contradicts constrained 2-anonymity w.r.t. *Location* attribute's maximal allowed generalization. On the other hand, the second microdata set, $\mathcal{MM}_2$, satisfies constrained 2-anonymity: every *QI*-cluster consists of at least 2 tuples, and none of the *Location* initial attribute's values are generalized beyond its *MAGVal*.

4 *GreedyCKA* - An Algorithm for Constrained *K*-Anonymization

In this section we assume that the initial microdata set $\mathcal{IM}$, the generalization boundaries for its quasi-identifier attributes, expressed as *MAGVal*s in their corresponding hierarchies, and the k value (as in k-anonymity) are given. First, we will describe a method to decide if $\mathcal{IM}$ can be masked to comply with constrained k-anonymity using generalization only, and second, we will introduce an algorithm for achieving constrained k-anonymity.

Our approach to constrained k-anonymization partially follows an idea found in [1] and [2], which consists in modeling and solving k-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). However, it is not always possible to mask an initial microdata to satisfy constrained k-anonymity only by generalization. Sometimes a solution to constrained k-anonymization has to combine generalization with suppression. In this case, our algorithm suppresses the *minimal* set of tuples

from IM such that is possible to build a constrained k-anonymous masked microdata for the remaining tuples.

The constrained k-anonymization by clustering problem can be formally stated as follows.

Definition 7. (***Constrained K-Anonymization by Clustering Problem***): Given a microdata IM, the ***constrained k-anonymization by clustering problem*** for IM is to find a partition $S = \{cl_1, cl_2, \ldots, cl_v, cl_{v+1}\}$ of IM, where $cl_j \subseteq IM$, j=1..v+1, are called clusters and: $\bigcup_{j=1}^{v} cl_j = IM - cl_{v+1}$; $cl_i \bigcap cl_j = \emptyset$, $i, j = 1..v+1$, $i \neq j$; $|cl_j| \geq k$, j=1..v ; and a cost measure is optimized. The cluster cl_{v+1} is formed of all the tuples in IM that have to be suppressed in MM, and the tuples within every cluster cl_j, j=1..v will be generalized (their quasi-identifier attributes) in MM to common values.

The generalization information of a cluster, which is introduced next, represents the minimal covering "tuple" for that cluster. Since in this paper we use predefined value generalization hierarchies for both categorical and numerical attributes, we do not have to consider a definition that distinguishes between these two types of attributes [21].

Definition 8. (***Generalization Information***): Let $cl = \{r_1, r_2, \ldots, r_u\}$ be a cluster of tuples selected from IM, $QI = \{Q_1, Q_2, \ldots, Q_s\}$ be the set of quasi-identifier attributes. The ***generalization information of cl*** w.r.t. 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 H_{Q_j} of $\{r_1[Q_j], \ldots, r_u[Q_j]\}$.

For the cluster cl, its generalization information $gen(cl)$ is the tuple having as value for each quasi-identifier attribute the most specific common generalized value for all that attribute values from *cl's* tuples. In the corresponding MM, each tuple from the cluster cl will have its quasi-identifier attributes values replaced by $gen(cl)$.

To decide whether an initial microdata can be masked to satisfy constrained k-anonymity property using generalization only, we introduce several properties. These properties will also allow us, in case that constrained k-anonymity cannot be achieved using generalization only, to select the tuples that must be suppressed.

Property 1. Let IM be a microdata set and cl a cluster of tuples from IM. If cl contains two tuples r_i and r_j such that $MAGVal(r_i[Q]) \neq MAGVal(r_j[Q])$, where Q is a quasi-identifier attribute, then the generalization of the tuples from cl to $gen(cl)$ will create at least one constraint violation.

Proof. Assume that there are two tuples r_i and r_j within cl such that $MAGVal(v_i) \neq MAGVal(v_j)$, where $v_i = r_i[Q]$ and $v_j = r_j[Q]$, $v_i, v_j \in leaves(H_Q)$. Let a be a value within H_Q that is the first common ancestor for $MAGVal(v_i)$ and $MAGVal(v_j)$. Depending on how $MAGVal(v_i)$ and $MAGVal(v_j)$ are located relatively to one another in the Q's value generalization hierarchy, a can be one of them, or a value on a superior tree level. In any case, a will be different from, and an ancestor for at least one of $MAGVal(v_i)$ or $MAGVal(v_j)$. This is a consequence of the fact that $MAGVal(v_i) \neq MAGVal(v_j)$: 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 v_i and v_j will be generalized further than its maximal allowed generalization value, leading to a constraint violation. // q.e.d.

Property 1 restricts the possible solutions of the constrained anonymization by clustering problem to those partitions S of IM for which every cluster to be generalized doesn't show any constraint violation w.r.t. each of the quasi-identifier attributes. The following definition introduces a masked microdata that will help us to express when the IM can be transformed to satisfy constrained 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, v, in IM is generalized to $MAGVal(v)$.

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

Proof. Assume that MAM is not k-anonymous, and there is a masked microdata MM that satisfies constrained k-anonymity. This means that every QI-cluster from MM has at least k elements 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, .., v). Because MM satisfies constrained k-anonymity, the generalization of 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 $MAGVal$s. As a consequence, by generalizing all quasi-identifier attributes values to their corresponding $MAGVal$s (this is the procedure to create the MAM microdata) all entities from the cluster cl_i (for all i = 1, .., v) will be contained within the same QI-cluster. This

means that each *QI*-cluster in $\mathcal{MAM}$ contains one or more *QI*-clusters from $\mathcal{MM}$ and its size will, then, be at least *k*. In conclusion, $\mathcal{MAM}$ is *k*-anonymous, which is a contradiction with our initial assumption. // q.e.d.

Property 3. If $\mathcal{MAM}$ satisfies *k*-anonymity then $\mathcal{MAM}$ satisfies the constrained *k*-anonymity property.

Proof. This follows from the definition of $\mathcal{MAM}$.

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

Proof. "If": If $\mathcal{MAM}$ satisfies *k*-anonymity, then based on Property 3, $\mathcal{MAM}$ is also constrained *k*-anonymous, and $\mathcal{IM}$ can be masked to $\mathcal{MAM}$ (in the worst case – or even to a less generalized masked microdata) to comply with constrained *k*-anonymity.

"Only If": If $\mathcal{MAM}$ does not satisfy *k*-anonymity, then based on Property 2, any masked microdata obtained by applying generalization only to $\mathcal{IM}$ will not satisfy constrained *k*-anonymity. // q.e.d.

Now we have all the tools required to check whether an initial microdata $\mathcal{IM}$ can be masked to satisfy the constrained *k*-anonymity property using generalization only. We follow the next two steps:

- Compute $\mathcal{MAM}$ for $\mathcal{IM}$. This is done by replacing each quasi-identifier attribute value with its corresponding *MAGVal*.
- If all *QI*-clusters from $\mathcal{MAM}$ have at least *k* entities than the $\mathcal{IM}$ can be masked to satisfy constrained *k*-anonymity.

It is very likely that there are some *QI*-clusters in $\mathcal{MAM}$ with size less than *k*. We use the notation $\mathcal{OUT}$ to represent all entities from these *QI*-clusters (for simplicity we use the same notation to refer to entities from both $\mathcal{IM}$ and $\mathcal{MAM}$). Unfortunately, the entities from $\mathcal{OUT}$ cannot be *k*-anonymized while preserving the constraint condition, as shown by the Property 6. For a given $\mathcal{IM}$ with its corresponding $\mathcal{MAM}$ and $\mathcal{OUT}$ sets the following two properties hold:

Property 5. $\mathcal{IM} \setminus \mathcal{OUT}$ can be masked using generalization only to comply with constrained *k*-anonymity.

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

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

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

The Properties 5 and 6 show that $\mathcal{OUT}$ is the minimal tuple set that must be suppressed from $\mathcal{IM}$ such that the remaining set could be constrained *k*-anonymized. To compute a constrained *k*-anonymous masked microdata using minimum suppression and generalization only we follow an idea found in [1] and [2], which consists in modeling and solving *k*-anonymization as a clustering problem. First, we suppress all tuples from the $\mathcal{OUT}$ set. Next, we create all *QI*-clusters in the maximum allowed microdata for $\mathcal{IM} \setminus \mathcal{OUT}$. Last, each such cluster will be divided further, if possible, using the clustering approach from [1, 2], into several clusters, all with size greater than or equal to *k*. This approach uses a greedy technique that tries to optimize an information loss (*IL*) measure. The information loss measure we use in our algorithm implementation was introduced in [2]. We present it in Definitions 10 and 11. Note that this *IL* definition assumes that value generalization hierarchies are predefined for all quasi-identifier attributes.

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

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

where:

- $|cl|$ denotes the cluster *cl* cardinality;
- $\Lambda(w)$, $w \in H_{Qj}$ is the subhierarchy of H_{Qj} rooted in *w*;
- $height(H_{Qj})$ is the height of the tree hierarchy H_{Qj}.

Definition 11. (***Total Information Loss***)**:** ***Total information loss*** for a partition $\mathcal{S}$ of the initial microdata set is the sum of the information loss measure for all clusters in $\mathcal{S}$.

In is worth noting that, for the constrained k-anonymization by clustering problem, 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: $IL(cl_{v+1}) = |cl_{v+1}| \cdot n$, where n is the number of quasi-identifier attributes. When performing experiments to compare the quality of constrained k-anonymous microdata and k-anonymous microdata, produced for the same IM, the information loss of the constrained k-anonymous solution includes the information loss caused by the suppressed cluster as well, and not only for the generalized clusters. More than that, for every suppressed tuple we consider the maximum information loss that it can cause when it is masked. This way, the quality of the constrained k-anonymous solutions will not be biased because of a favored way of computing information loss for the suppressed tuples.

The two-stage constrained k-anonymization algorithm called *GreedyCKA* is depicted in Figure 5.

We present below the pseudocode of the *GreedyCKA* Algorithm:

```
Algorithm GreedyCKA is
Input   IM – initial microdata;
        k – as in k-anonymity;
Output  S={cl1,cl2,… clv,clv+1} – a solution for
        the constrained k-anonymization by
        clustering problem for IM;
 Compute MAM and OUT;
 S = ∅;
 For each QI-cluster from MAM \ OUT, cl,
 {
   // By cl we refer to the entities from IM
   // that are clustered together in MAM.
   S' = Greedy_k-member_Clustering(cl, k); //[2]
   S = S ∪ S';
 }
 v = |S|;
 clv+1 = OUT;
End GreedyCKA;
```

This idea of dividing IM into clusters based on common *MAGVal*s of the quasi-identifiers can be employed for other privacy models as well, not only for k-anonymity. For instance, if we use an algorithm that creates a p-sensitive k-anonymous masked microdata [20], such as *EnhancedPKClustering* [22], we just need to execute that algorithm instead of *Greedy_k-member_Clustering*, for each *QI*-cluster from $MAM \setminus OUT$. The obtained masked microdata will be p-sensitive k-anonymous and will satisfy the generalization boundaries. We can define this new privacy model as ***constrained p-sensitive k-anonymity***. Using similar modifications in the *GreedyCKA* algorithm, we can introduce constrained versions of other privacy models such as: ***constrained l-diversity*** [12], ***constrained t-closeness*** [10], etc. and generate their corresponding masked microdata sets.

5 Experimental Results

In this section we compare the *GreedyCKA* and *Greedy_k-member_Clustering* [2] algorithms with respect to: the quality of the results they produce measured against the information loss measure; the algorithms' efficiency as expressed by their running time; the number of constraint violation that k-anonymous masked microdata produced by *Greedy_k-member_Clustering* have; and the suppression amount performed by *GreedyCKA* in order to produce constrained k-anonymous masked microdata in presence of different constraint sets.

The two algorithms were implemented in Java; tests were executed on a dual CPU machine with 3.00 GHz and 1 GB of RAM, running Windows 2003 Server.

A set of experiments were performed for an IM consisting of 10,000 tuples randomly selected from the *Adult* dataset from the UC Irvine Machine Learning Repository [15]. In all the experiments, we considered a set of eight quasi-identifier attributes: *education-num*, *workclass*, *marital-status*, *occupation*, *race*, *sex*, *age*, and *native-country*.

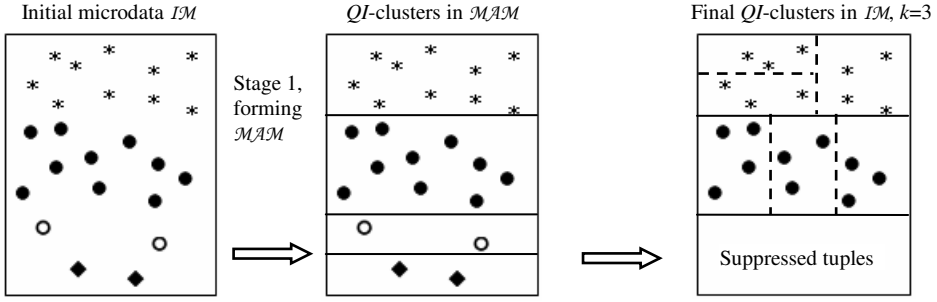

Figure 5: The two-stage process in creating constrained k-anonymous masked microdata.

The *GreedyCKA* and *Greedy_k-member_Clustering* algorithms were applied to this microdata set, for different k values, from k=2 to k=10. Two different generalization constraint sets were successively considered for every k value. First, only the *native-country* attribute's values were subject to generalization constraints, as depicted in Figure 6. Second, both *native-country* and *age* had generalization boundaries; the value generalization hierarchy and the maximum allowed generalization values for the *age* attribute are illustrated in Figure 7. In Figures 6 and 7, the *MAGVal*s are depicted as bold and delimited by * characters. Of course, *Greedy_k-member_Clustering* proceeded without taking into consideration the generalization boundaries, as it is a "simple", unconstrained k-anonymization algorithm. This is why the masked microdata it produces will generally contain numerous constraint violations. On the other side, the k-anonymization process of *GreedyCKA* is conducted in respect to the specified generalization boundaries; this is why the masked microdata produced by *GreedyCKA* is free of constraint violations.

The quasi-identifier attributes without generalization boundaries have the following heights for their corresponding value generalization hierarchies: *education-num* – 4, *workclass* – 1, *marital-status* – 2, *occupation* – 1, *race* – 1, and *sex* – 1.

However, masking microdata to comply with the more restrictive constrained k-anonymity model sometimes comes with a price. As the experiments show, it is possible to lose more of the intrinsic microdata information when masking it to satisfy constrained k-anonymity than when masking it to satisfy k-anonymity only. Figure 8 presents comparatively the information loss measure for the masked microdata created by *GreedyCKA* and *Greedy_k-member_Clustering*, with the two different constraint sets and for k values in the range 2-10.

As expected, the information loss value is generally greater when constraints are considered in the k-anonymization process. Exceptions may however occur. For example, *GreedyCKA* obtained better results then *Greedy_k-member_Clustering* for k = 8, 9 and 10, when only *native_country* was constrained. The information lost is influenced, of course, by the constraint requirements and by the microdata distribution w.r.t. the constrained attributes. When more quasi-identifiers have generalization boundaries or more restrictive generalization boundaries, the information lost in the constrained k-anonymization process will generally increase.

Regarding the running time, we can state that *GreedyCKA* will always be more efficient than *Greedy_k-member_Clustering*. The explanation for this fact is that, when generalization boundaries are imposed, they will cause the initial microdata to be divided in several subsets (the *QI*-clusters of $\mathcal{MAM}$), on which *Greedy_k-member_Clustering* will be afterwards applied. *Greedy_k-member_Clustering* has an $O(n^2)$ complexity, and applying it on smaller microdata subsets will reduce the processing time. More constraints and *QI*-clusters exist in $\mathcal{MAM}$, more significant is the reduction of the processing time for microdata masking (see Figure 9).

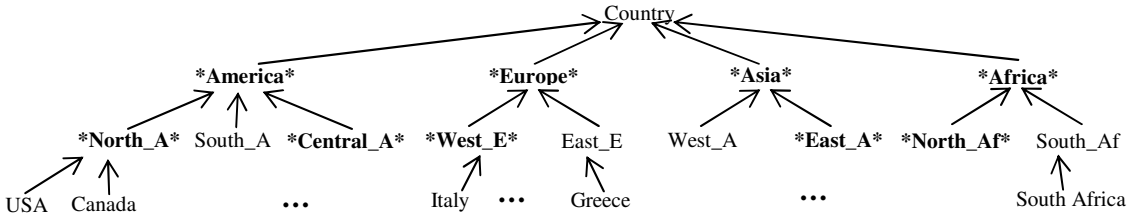

Figure 6: *MAGVals* for the quasi-identifier attribute *Country*.

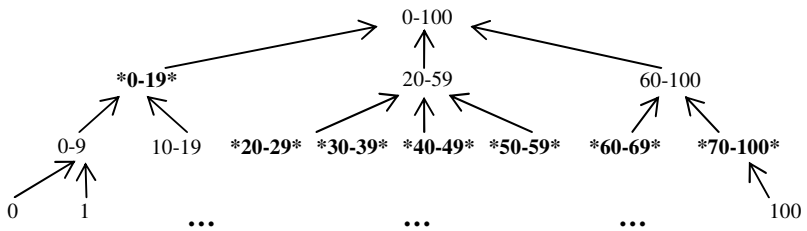

Figure 7: *MAGVals* for the quasi-identifier attribute *Age*.

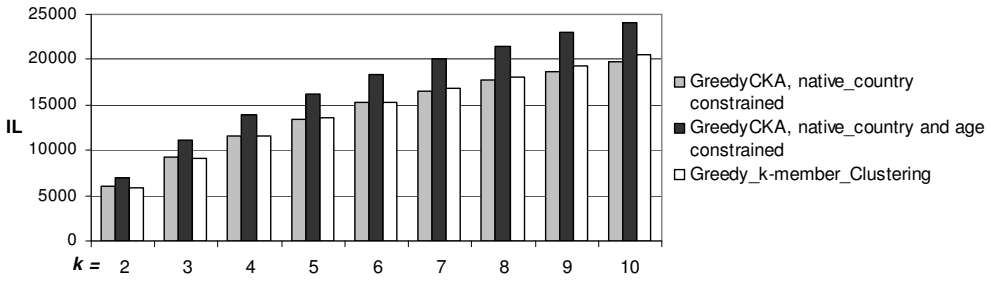

Figure 8: Information Loss (*IL*) for *GreedyCKA* and *Greedy_k-member_Clustering*.

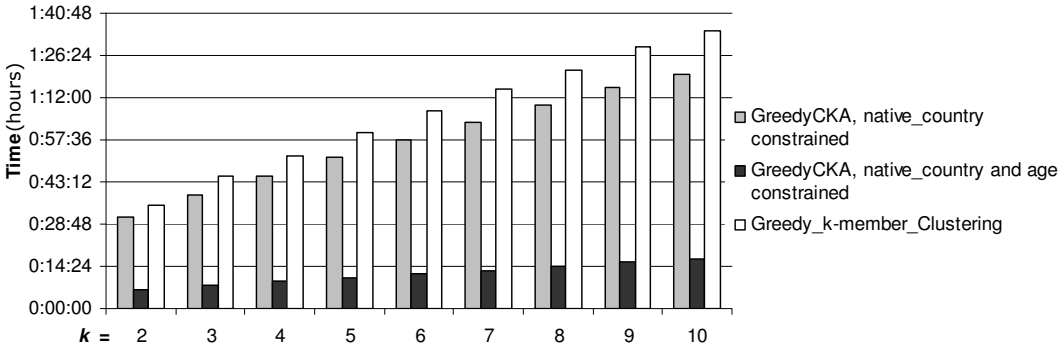

Figure 9: Running Time for *GreedyCKA* and *Greedy_k-member_Clustering*.

As pointed out, when *Greedy_k-member_Clustering* is applied to *k*-anonymize *IM*, the resulting masked microdata usually contains numerous constraint violations. Table 1 reports the number of constraint violations in the outcome of the *Greedy_k-member_Clustering* unconstrained *k*-anonymization algorithm, for two maximal generalization requirement sets.

k	No of constraint violations for 1 constrained attribute – **native_country**	No of constraint violations for 2 constrained attributes – **native_country, age**
2	605	2209
3	991	3824
4	1377	5297
5	1657	6163
6	1906	6964
7	2198	7743
8	2354	8417
9	2550	8931
10	2728	9549

Table 1: Constraint violations in *Greedy_k-member_Clustering*

k	2	3	4	5	6	7	8	9	10
No of suppressed tuples for 1 constrained attribute – **native_country**	0	0	0	0	0	0	0	0	0
No of suppressed tuples for 2 constrained attributes– **native_country, age**	5	15	24	28	48	60	81	97	106

Table 2: Number of tuples suppressed by *GreedyCKA*

Table 2 shows the number of tuples suppressed by *GreedyCKA*, while masking the initial microdata.

All in all, our experiments proved that constrained *k*-anonymous masked microdata can be achieved without sacrificing the data quality to a significant extent, when compared to a corresponding *k*-anonymous unconstrained masked microdata.

While the constrained *k*-anonymity model responds to a necessity in real-life applications, the existing *k*-anonymization algorithms are not able to build masked microdata that comply with it. In this context, *GreedyCKA* takes optimal suppression decisions, based on the proved properties of the new model (Properties 5 and 6), and builds high-quality constrained *k*-anonymous masked microdata.

6 Conclusions and Future Work

In this paper we defined a new privacy model, called constrained *k*-anonymity, which takes into consideration generalization boundaries imposed by the data owner for quasi-identifier attributes. Based on the model properties, an efficient algorithm to generate a masked microdata to comply with constrained *k*-anonymity property was introduced. Our experiments showed that the proposed algorithm obtains comparable information loss values with *Greedy_k-member_Clustering* algorithm, while the masked microdata sets obtained by the latter have many constraint violations.

In this paper we used predefined hierarchies for all quasi-identifier attributes. As future work we plan to extend this concept further for numerical attributes. We plan to provide a technique to dynamically determine for each numerical quasi-identifier value, its maximal allowed generalization, based on that attribute's values in the analyzed microdata and a minimal user input.

We also pointed out that the constraint *k*-anonymity property and even our proposed algorithm, *GreedyCKA*, can be extended to other privacy models (models such as constrained *l*-diversity, constrained (α, *k*)-anonymity, constrained *p*-sensitive *k*-anonymity, etc. can be easily defined). Finding specific properties for these enhanced privacy models, and developing improved algorithms to generate masked microdata to comply with such models are subject of future work.

Acknowledgments

This work was partially supported by the CNCSIS (Romanian National University Research Council) grant PNCDI-PN II, IDEI 550/2007.

References

[1] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu, *Achieving Anonymity via Clustering*, in Proc. of the ACM PODS (2006), pp. 153–162.

[2] J.W. Byun, A. Kamra, E. Bertino, and N. Li, *Efficient k-Anonymization using Clustering Techniques*, in Proc. Of DASFAA (2006), pp. 188–200.

[3] A. Campan, T. M. Truta, *Extended P-Sensitive K-Anonymity*, Studia Universitatis Babes-Bolyai, Informatica, Vol. 51, No. 2 (2006), pp. 19–30.

[4] B. C. M. Fung, K. Wang, and P. S. Yu, *Anonymizing classification data for privacy preservation*, IEEE Transactions on Knowledge and Data Engineering, Vol. 19, No. 5 (2007), pp. 711–725.

[5] G. Ghinita, K. Karras, P. Kalinis, and N. Mamoulis, *Fast Data Anonymization with Low Information Loss*, in Proc. of VLDB (2007), pp. 758–769.

[6] HIPAA, *Health Insurance Portability and Accountability Act*, www.hhs.gov/ocr/hipaa, 2002.

[7] V. Iyengar, *Transforming Data to Satisfy Privacy Constraints*, in Proc. of the ACM SIGKDD (2002), pp. 279–288.

[8] K. LeFevre, D. DeWitt, and R. Ramakrishnan, *Incognito: Efficient Full-Domain K-Anonymity*, in Proc. of the ACM SIGMOD (2005), pp. 49–60.

[9] K. LeFevre, D. DeWitt, and R. Ramakrishnan, *Mondrian Multidimensional K-Anonymity*, in Proc. of the IEEE ICDE (2006), pp. 25.

[10] N. Li, T. Li, and S. Venkatasubramanian, *T-Closeness: Privacy Beyond k-Anonymity and l-Diversity*, in Proc. of the IEEE ICDE (2007), pp. 106–115.

[11] M. Lunacek, D. Whitley, and I. Ray, *A Crossover Operator for the k-Anonymity Problem*, in Proc. of the GECCO (2006) pp. 1713–1720.

[12] A. Machanavajjhala, J. Gehrke, and D. Kifer, *L-Diversity: Privacy beyond K-Anonymity*, in Proc. of the IEEE ICDE (2006), pp. 24.

[13] D. J. Martin, D. Kifer, A. Machanavajjhala, and J. Gehrke, *Worst-Case Background Knowledge for Privacy-Preserving Data Publishing*, Proc. of the IEEE ICDE (2007), pp. 126–135.

[14] MSNBC, *Privacy Lost*, www.msnbc.msn.com/id/ 15157222 , 2006.

[15] D. J. Newman, S. Hettich, C. L. Blake, and C. J. Merz, *UCI Repository of Machine Learning Databases*, www. ics.uci.edu/~mlearn/MLRepository.html, 1998.

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

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

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

[19] T. M. Truta, F. Fotouhi, and D. Barth-Jones, *Privacy and Confidentiality Management for the Microaggregation Disclosure Control Method*, in Proc. of the PES Workshop, with ACM CCS (2003), pp. 21–30.

[20] T. M. Truta, V. Bindu, *Privacy Protection: P-Sensitive K-Anonymity Property*, in Proc. of the PDM Workshop, with IEEE ICDE (2006), pp. 94.

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

[22] T. M. Truta, A. Campan, P. Meyer, *Generating Microdata with P-Sensitive K-Anonymity Property*, in Proc. of the SDM Workshop, with VLDB (2007), pp. 124–141.

[23] W. Winkler, *Matching and Record Linkage*, Business Survey Methods, Wiley (1995), pp. 374–403.

[24] R. C. W. Wong, J. Li, A. W. C. Fu, and K. Wang, *(α, k)-Anonymity: An Enhanced k-Anonymity Model for Privacy-Preserving Data Publishing*, in Proc. of the ACM SIGKDD (2006), pp. 754–759.

[25] R. C. W. Wong, J. Li, A. W. C. Fu, and J. Pei, *Minimality Attack in Privacy-Preserving Data Publishing*, in Proc. of the VLDB (2007), pp. 543–554.

[26] X. Xiao, Y. Tao, *Personalized Privacy Preservation*, in Proc. of the ACM SIGMOD (2006), pp. 229–240.

[27] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A. Fu, *Utility-Based Anonymization Using Local Recoding*, in Proc. of ACM SIGKDD (2006), pp. 785–790.

[28] Q. Zhang, N. Koudas, D. Srivastava, and T. Yu, *Aggregate Query Answering on Anonymized Tables*, in Proc. Of the IEEE ICDE (2007), pp. 116–125.