STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume LI, Number 2, 2006

# EXTENDED *P*-SENSITIVE *K*-ANONYMITY

ALINA CAMPAN AND TRAIAN MARIUS TRUTA

Abstract. In this paper we introduce a new privacy protection property, called *extended p-sensitive k-anonymity*, which is an extension of the $p$-sensitive $k$-anonymity property [16]. The new property is aware of confidential attributes hierarchies and of the existence of protected not ground-level confidential attributes values, situation not considered by previous work done in this direction. We describe our model and indicate an algorithm for enforcing extended $p$-sensitive $k$-anonymity to masked microdata.

**Keywords:** privacy protection, anonymity, generalization.

## 1. Introduction

To protect the privacy of individuals in the present digitized world became an increasingly difficult task. Large amounts of *microdata* (datasets where each tuple belongs to an individual entity) are collected by different agencies. Some of these microdata need to be released, for various purposes, to other parties. Obviously, direct identifying information such as *SSN*, *Name* is eliminated from the microdata before releasing it, for privacy protection. But even modified this way, the datasets could still present vulnerabilities that can be exploited by intruders, i.e. persons whose goals are to identify specific individuals and to use the confidential information they discover for malicious purposes. More elaborated techniques are needed in order to ensure a reliable and controlled privacy protection when microdata are released.

In recent years, the use and the disclosure of confidential information was subject to privacy regulations promulgated in different domains [4, 8, 7]. All these regulations, together with the necessity of collecting personal information, have fed the interest in privacy research.

Techniques to avoid the disclosure of confidential information exist in the literature [1, 17]. Among them, the $k$-anonymity property required for the released

Received by the editors: 20/09/2006.

2000 *Mathematics Subject Classification.* 68P15.

1998 *CR Categories and Descriptors.* 68P15 [**Computer science**]: Theory of data – *Database theory.*

microdata (a.k.a. masked microdata) was recently introduced [13, 14] and extensively studied [3, 5, 10, 16]. This property requires that in the released microdata every tuple will be indistinguishable from at least ($k$-1) other tuples with respect to a subset of attributes called quasi-identifier attributes or key attributes.

Recent results have showed that $k$-anonymity fails to protect the privacy of individuals in all situations [16]. Two similar models called $p$-sensitive $k$-anonymity [16] and $l$-diversity [11] were proposed in the literature in order to deal with the problems of the $k$-anonymity model. The $p$-sensitive $k$-anonymity property requires, in addition to $k$-anonymity, that for each group of tuples with the identical combination of quasi-identifier attributes values, the number of distinct values for each confidential attribute (attribute which values must be protected) must be at least $p$ within the same group.

However, depending on the nature of the confidential attributes, even the $p$-sensitivity property still permits the information to be disclosed. We identify, in this paper, situations when $p$-sensitivity property is not enough for privacy protection and we propose a solution to overcome the identified problem: the extended $p$-sensitive $k$-anonymity model and an algorithm to enforce this property.

## 2. Concepts and Notations

Let $IM$ be the initial microdata and $IM$ be the released (a.k.a. masked) microdata. $IM$ consists in a set of tuples over an attribute set. The attributes characterizing microdata are classified into the following three categories:

- $I_1, I_2, \ldots, I_m$ are *identifier* attributes such as *Name* and *SSN* that can be used to identify a record. These attributes are present only in the initial microdata because they express information which can lead to a specific entity.
- $K_1, K_2, \ldots, K_n$ are *key* or *quasi-identifier* attributes such as *ZipCode* and *Age* that may be known by an intruder. Quasi-identifier attributes are present in the masked microdata as well as in the initial microdata.
- $S_1, S_2, \ldots, S_r$ are *sensitive* or *confidential* attributes such as *PrincipalDiagnosis* and *ICD9Code* that are assumed to be unknown to an intruder. Confidential attributes are present in the masked microdata as well as in the initial microdata.

While the identifier attributes are removed from the released microdata, the quasi-identifier and confidential attributes are usually kept in the masked microdata and released to the researchers.

A general assumption, as noted, is that the values for the confidential attributes are not available from any external source. This assumption guarantees that an intruder can not use the confidential attributes values to increase his/her chances of disclosure. Unfortunately, an intruder may use record linkage techniques [18] 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.

**Definition 1.** (***k-anonymity property***): The ***k-anonymity property*** for a masked microdata ($MM$) is satisfied if every combination of quasi-identifier attribute values in $MM$ occurs $k$ or more times.

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

To achieve $k$-anonymity, existing $k$-anonymization algorithms generally proceed by using generalization and suppression [13, 15]. Generalization of the quasi-identifier attributes is used widely for $k$-anonymization. It consists in replacing the actual value of an attribute with a less specific, more general value that is faithful to the original [15]. Generalization is either based on predefined (static) domain and value generalization hierarchies [15], or is conducted using a hierarchy-free model [10].

The $k$-anonymity property ensures protection against identity disclosure, i.e. the identification of an entity (person, institution). However, as we will show next, it does not protect the data against attribute disclosure, which occurs when the intruder finds something new about a target entity. The two disclosure types are independent. None of them does imply the other.

Consider the masked microdata example below, where the set of quasi-identifier attributes is composed of *Age*, *ZipCode* and *Gender*, and *Illness* is the sensitive attribute:

Table 1. Patient masked microdata satisfying 2-anonymity

| **Tuples** | **Age** | **ZipCode** | **Gender** | **Illness** |
|---|---|---|---|---|
| $r_1$ | 50-60 | 43102 | Male | Colon Cancer |
| $r_2$ | 30-40 | 43102 | Female | Breast Cancer |
| $r_3$ | 30-40 | 43102 | Female | HIV |
| $r_4$ | 20-30 | 43102 | Male | Diabetes |
| $r_5$ | 20-30 | 43102 | Male | Diabetes |
| $r_1$ | 50-60 | 43102 | Male | Heart Disease |

Identity disclosure does not happen in this masked microdata, as its construction guarantees that for every existing combination of values for *Age*, *ZipCode* and *Gender* there are at least two tuples that have the respective combination of values. However, assuming that external information in Table 2 below is available, attribute disclosure can take place. If the intruder knows that in the masked microdata the *Age* attribute was generalized to multiples of 10, he can deduce that both Sam and Eric have Diabetes, even he doesn't know which tuple, $r_4$ or $r_5$, corresponds to what person. This example shows that $k$-anonymity fails to protect sometimes against attribute disclosure, even if it protects from identity disclosure.

TABLE 2. External information for Patient example

| **Name** | **Age** | **Gender** | **ZipCode** |
|---|---|---|---|
| Sam | 29 | Male | 43102 |
| Gloria | 38 | Female | 43102 |
| Adam | 51 | Male | 43102 |
| Eric | 29 | Male | 43102 |
| Dana | 34 | Female | 43102 |
| Don | 51 | Male | 43102 |

For dealing with this flaw in privacy, another model, called $p$-sensitive $k$-anonymity was introduced in [16]. A similar privacy model, called $l$-diversity, is described in [11].

**Definition 2.** (***p-sensitive k-anonymity property***): The masked microdata ($MM$) satisfies ***p-sensitive k-anonymity property*** if it satisfies $k$-anonymity and for each group of tuples with the identical combination of key attribute values that exists in $MM$, the number of distinct attributes for each confidential attribute is at least $p$ within the same group.

Sometimes, similar to the quasi-identifier attributes, the domain of the sensitive attributes, especially the categorical ones, can also be organized according to some hierarchies. For example, in medical datasets, the *Illness* attribute has values as specified by the ICD9 codes (see Figure 2). The different types of diseases are organized in a tree hierarchy of values. The attribute values are very specific, for example they can represent different types of cancer, which are all descendants of cancer value. The initial microdata contain as values for the *Illness* attribute values from the lowest level of the hierarchy (i.e. from the leaf nodes). In these conditions, the data owner can be interested in protecting not only these most specific values, but also information found at higher levels. For example, the information that a person has cancer needs to be protected, regardless of the cancer type she has. If $p$-sensitive $k$-anonymity property is enforced for masked microdata, it is possible that in a group with $p$ distinct *Illness* attribute values, all of them to be descendants of the cancer node in the corresponding hierarchy. To avoid such situations, we introduce the concept of extended $p$-sensitive $k$-anonymity, which is aware of the existence of protected values not only at the ground level.

## 3. Extended $p$-sensitive $k$-anonymity Property

Let $S$ be a categorical confidential attribute we want to protect against attribute disclosure. $S$ has associated predefined (static) domain and value generalization hierarchies [15]. $HD_S$ is the domain generalization hierarchy of attribute $S$. The values from different domains of this hierarchy $HD_S$ are represented in a tree $HV_S$ called value generalization hierarchy. We illustrate domain and value generalization hierarchy in Figure 1 for attributes *ZipCode* and *Gender*, which are quasi-identifier attributes.

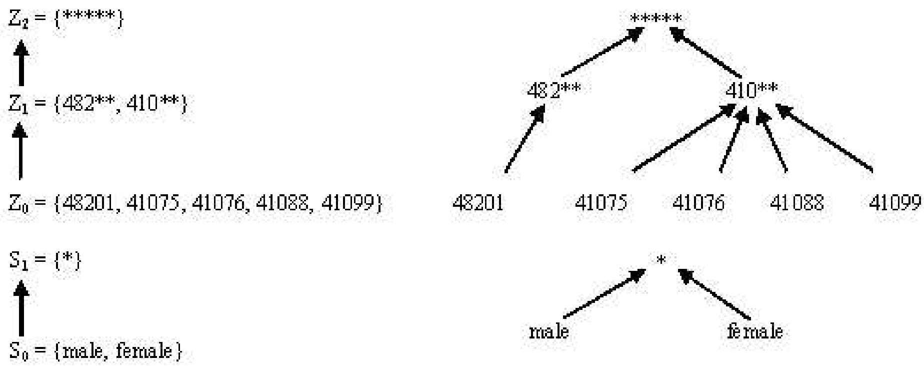


Figure 1. Examples of domain and value generalization hierarchies

Figure 2 shows a part of the ICD9 value generalization hierarchy.

Some zones of a value generalization hierarchy $HV_S$, associated to the sensitive attribute $S$, need to be protected.

Unlike the quasi-identifier attributes, the values of a sensitive attribute cannot be generalized in the masked microdata for protection, because this would affect the quality of the released data w.r.t. subsequent tasks that will be performed on it, such as data mining tasks.

The protection will be achieved by enforcing $k$-anonymity (for identity disclosure protection) while ensuring the extended $p$-sensitivity (for attribute disclosure protection). The heterogeneity of the confidential attributes values in each of the groups formed by $k$-anonymizing the data is to be achieved not only at the ground values level, but for all the values declared protected in $HV_S$. The data owner has to mark (declare) which are the protected "zones" in a confidential attribute hierarchy. In Figure 2, the protected values in the value generalization hierarchy of attribute *Illness* are bordered. We require that all the descendants of a protected value to also be protected. In other words, if an internal node of a value generalization hierarchy is protected, the entire subtree rooted in that node needs to be

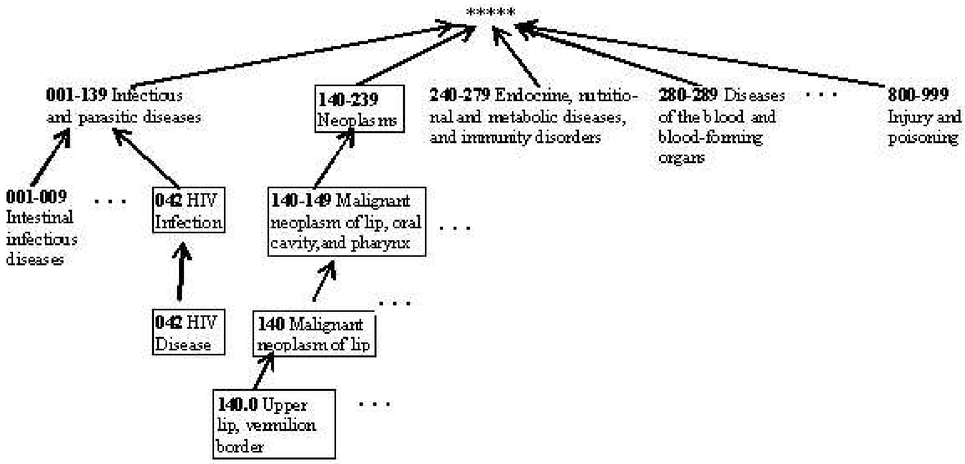


FIGURE 2. ICD9 disease hierarchy and codes

protected. All values at the ground level are considered to be protected. The semantics of a node (its value) being protected is as follows: if extended $p$-sensitivity is enforced for a microdata w.r.t. the confidential attribute $S$, this means that each group of tuples with the identical combination of quasi-identifier attributes values contains at least $p$ distinct values for $S$ that respect the condition that, any two of them are not descendants of a common protected value (i.e. any two of these values do not have a common protected ancestor). For example, if Neoplasms is a protected value, no group will contain only descendant values of Neoplasms, even if there are $p$ distinct such values in that group. So, every group containing descendant values of Neoplasms will also contain at least $p$ - 1 different values that are not descendants of Neoplasms. Of course, for these values also functions the same condition. We will refer to the property enounced here informally as extended $p$-sensitive $k$-anonymity. To define the extended $p$-sensitive $k$-anonymity property we need to introduce several other concepts.

**Requirements**: Let $S$ be a confidential attribute and $HV_S$ its value generalization hierarchy. The following two requirements must be met by the protected values in $HV_S$:

- All ground values in $HV_S$ are protected.
- All the descendants of a protected internal value in $HV_S$ are protected.

**Definition 3.** A protected value in the value generalization hierarchy $HV_S$ of a confidential attribute $S$ is called ***strong*** if none of its ascendants (including the root) is protected.

**Property 1.** A protected value is strong if its parent is not protected.

This property results from the definition of strong values and the first requirement imposed to $HV_S$.

**Definition 4.** We call ***protected subtree*** of a hierarchy $HV_S$ a subtree in $HV_S$ that has as root a strong protected value.

**Definition 5.** (***extended p-sensitive k-anonymity property***): The masked microdata ($MM$) satisfies ***extended p-sensitive k-anonymity property*** if it satisfies $k$-anonymity and for each group of tuples with the identical combination of key attribute values that exists in $MM$, the values of each confidential attribute $S$ within that group belong to at least $p$ different protected subtrees in $HV_S$.

Extended $p$-sensitive $k$-anonymity can not be enforced for any microdata set. We give next several necessary conditions that must be satisfied by a microdata set in order to be possible to enforce extended $p$-sensitive $k$-anonymity for it. These conditions are adapted from [16], where they were enounced w.r.t. the basic $p$-sensitive $k$-anonymity property.

**Condition 1.** $p$ must be less than or equal to $k$ (i.e. $p \leq k$).

Justification: In a group of $k$ tuples there can not be more than $k$ different values for a confidential attribute $S$.

**Condition 2.** The value generalization hierarchy $HV_S$ of every confidential attribute $S$ must contain at least $p$ different protected subtrees.

We use the following notations for a microdata $IM$:

- $n$ - the number of tuples in $IM$;
- $q$ - the number of confidential attributes in $IM$;
- $s_j$ - the number of distinct strong protected values in $HV_{S_j}$ that are ascendants of all the values that the confidential attribute $S_j$ has in $IM$, $1 \leq j \leq q$;
- $f_i^j$- the descending ordered frequency set for the confidential attribute $S_j$, $1 \leq i \leq s_j$, $1 \leq j \leq q$. The frequency set is computed after the confidential values in the microdata are generalized to their corresponding strong protected values;
- $cf_i^j$ - the cumulative descending ordered frequency set for the confidential attribute $S_j$, $1 \leq j \leq q$. The frequency set is computed after the confidential values in the microdata are generalized to their corresponding strong protected values;
- $cf_i = max_{j=1,q}(cf_i^j)$, $1 \leq i \leq min_{j=1,q}(s_j)$.

**Condition 3.** The maximum allowed number of combinations of quasi-identifier attribute values in the masked microdata $MM$ is $min_{i=1,p-1}\frac{n-cf_{p-i}}{i}$.

The proof of this property for basic $p$-sensitive $k$-anonymity can be found in [16]. For extended $p$-sensitivity, the confidential attributes values are first generalized in the initial microdata, to their strong ancestors, and then the property for basic $p$-sensitivity is true for the resulted dataset.

## 4. Enforcing Extended $p$-sensitive $k$-anonymity Property to Microdata

At a closer look, extended $p$-sensitive $k$-anonymity for a microdata is equivalent to $p$-sensitive $k$-anonymity for the same microdata where the confidential attributes values are generalized to their first protected ancestor, starting from the hierarchy root (their strong ancestor). Consequently, in order to enforce extended $p$-sensitive $k$-anonymity to a dataset, the following two-steps procedure can be applied:

- Each value of a confidential attribute is generalized (only temporarily) to its first protected ancestor (including itself), starting from the hierarchy root, i.e. to its strong ancestor.
- Any algorithm which can be used for $p$-sensitive $k$-anonymization is applied to the modified dataset. Such an algorithm is indicated in [16]. In the resulted masked microdata the original values of the confidential attributes are restored.

The dataset obtained following these steps respects the extended $p$-sensitive $k$-anonymity property.

## 5. Experimental Results

We performed a set of experiments to test how the existing $k$-anonymizing algorithms break the $p$-sensitivity and extended $p$-sensitivity properties. These experiments show that attribute disclosure can happen when only $k$-anonymity is enforced for microdata and, therefore, emphasize the need to protect the data against disclosure, beyond the $k$-anonymity.

In our experiments we used data based on the Adult database from the UC Irvine Machine Learning Repository [12]. This database has become the benchmark in data privacy field, being used by many researchers [10]. We considered *Age*, *Marital_Status*, *Race* and *Sex* from adult data as being the set of quasi-identifier attributes. The confidential attributes are *Pay*, *Capital_Gain*, *Capital_Loss* and *Tax_Amount*. The *Pay* attribute is considered to have two distinct values, $\leq$50K, >50K, and describes whether a person makes or not over 50K a year. The *Capital_Gain* attribute can have three distinct values (1000, 2000, 3000), *Capital_Loss* has four distinct values (1000, 2000, 3000, 4000), and *Tax_Amount* has ten distinct values (100, 200, $\ldots$, 1000). The *Tax_Amount* attribute is the only confidential attribute that has an associated generalization hierarchy with more than one level. The value generalization hierarchy is depicted in Figure 3, and the protected values are bordered, the strong protected values are bold bordered.

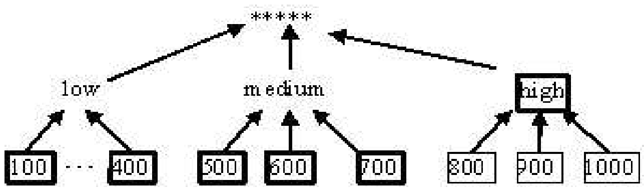


FIGURE 3. Value generalization hierarchy for *Tax_Amount*

We $k$-anonymized 400 records randomly chosen from adult database, for $k$=3 and $k$=5, using: the anonymization algorithm based on clustering which is described in [6]; the binary search algorithm presented in [13]. The quasi-identifier attributes were generalized w.r.t. the generalizations outlined in Table 3.

TABLE 3. Adult database quasi-identifier attributes generalization

| Attribute | First Generalization | Second Generalization | Third Generalization |
|---|---|---|---|
| Age | 10-years range | $\leq$50 and >50 groups | One group |
| Marital_Status | Single or Married | One group | - |
| Race | White, Black or Other | White or Other | One group |
| Sex | One group | - | - |

The produced masked microdata respect of course the requirements imposed by the $k$-anonymity property, but it contains several records that contradict the conditions in $p$-sensitive $k$-anonymity and in extended $p$-sensitive $k$-anonymity. Table 4 summarizes the results of our experiments: the number of tuples and the number of groups of tuples sharing common values for the quasi-identifier attributes that contradict the two properties. So, this experiment shows that for microdata masked to satisfy the $k$-anonymity property, disclosure channels still exist so that confidential attributes values can be inferred. $P$-sensitive $k$-anonymity property, basic or extended, need to be enforced to the microdata in order to avoid such disclosure situations. We used for $k$-anonymization two different algorithms, reported in [13], and respectively in [6].

Table 4. Attribute disclosures for a masked microdata set with $k$-anonymity property

| **$k$-anonymity with [13] algorithm** | **No of attribute disclosures w.r.t. $p$-sensitivity** |
|---|---|
| 2-anonymity | 6 |
| 3-anonymity | 2 |

| **$k$-anonymity with [6] algorithm** | ***Pay*** | | ***Capital_Gain*** | | ***Capital_Loss*** | | ***Tax_Paid*** | |
|---|---|---|---|---|---|---|---|---|
| 3-anonymity | 2-sensitivity disclosures | | | | | | | |
| | Tuples | Groups | Tuples | Groups | Tuples | Groups | Tuples | Groups |
| | 38 | 12 | 36 | 12 | 15 | 5 | 0 | 0 |
| 5-anonymity | 3-sensitivity disclosures | | | | | | | |
| | Tuples | Groups | Tuples | Groups | Tuples | Groups | Tuples | Groups |
| | - | - | 164 | 31 | 30 | 6 | 11 | 2 |

| **$k$-anonymity with [6] algorithm** | ***Tax_Paid* extended $p$-sensitivity disclosures** | |
|---|---|---|
| 3-anonymity | extended 2-sensitivity disclosures | |
| | Tuples | Groups |
| | 3 | 1 |
| 5-anonymity | extended 3-sensitivity disclosures | |
| | Tuples | Groups |
| | 11 | 2 |

## 6. Conclusions and Future Work

In this paper, we introduced a new privacy protection property, called extended $p$-sensitive $k$-anonymity, which is an extension of the $p$-sensitive $k$-anonymity property. Next, we presented three necessary conditions a masked microdata must satisfy in order to have extended $p$-sensitive $k$-anonymity property. Last, we indicated how an algorithm that generates $k$-anonymous microdata can be modified to enforce extended $p$-sensitive $k$-anonymity property. Our experiments showed that $p$-sensitive $k$-anonymity property, basic or extended, need to be enforced to the masked microdata in order to avoid attribute disclosure situations.

In future work, we will create masked microdata that satisfy extended $p$-sensitive $k$-anonymity using the existing algorithms for $k$-anonymity with the addition of the three necessary conditions, and we will compare the running time of these modified algorithms against the existing algorithms that search for $k$-anonymity only.

## References

[1] Adam N. R., Wortmann J. C. (1989), "Security Control Methods for Statistical Databases: A Comparative Study." ACM Computing Surveys, Vol. 21, No. 4, 515–556.

[2] Aggarwal G., Feder T., Kenthapadi K., Khuller S., Panigrahy R., Thomas D., and Zhu A. (2006), "Anonymizing Tables", Proceedings of the ACM PODS Conference, 153–162.

[3] Aggarwal G., Feder T., Kenthapadi K., Motwani R., Panigrahy R., Thomas D., and Zhu A. (2005), "Achieving Anonymity via Clustering", Proceedings of the 10th International Conference on Database Theory.

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

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

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

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

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

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

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

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

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

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

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

[15] 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, 571–588.

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

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

[18] Winkler W.E. (1994), "Advanced Methods for Record Linkage", Proceedings of the Section on Survey Research Methods, American Statistical Society, 467–472.

Babeş Bolyai University, Cluj Napoca, Romania
*E-mail address*: `alina@cs.ubbcluj.ro`

Northern Kentucky University, Highland Heights, USA
*E-mail address*: `trutat1@nku.edu`