User-Controlled Generalization Boundaries for *P*-Sensitive *K*-Anonymity

Alina Campan
Department of Computer Science
Northern Kentucky University
Highland Heights, KY 41099, USA
001-859-572-5776
campana1@nku.edu

Traian Marius Truta
Department of Computer Science
Northern Kentucky University
Highland Heights, KY 41099, USA
001-859-572-7551
trutat1@nku.edu

Nicholas Cooper
Department of Computer Science
Northern Kentucky University
Highland Heights, KY 41099, USA
001-859-572-6930
coopern1@nku.edu

ABSTRACT
Numerous privacy models based on the *k*-anonymity property have been introduced in the last few years. 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. 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 quality-wise with existing algorithms.

Categories and Subject Descriptors
K.4.1 [**Computers and Society**]**:** Public Policy Issues – *privacy*. I.5.3 [**Pattern Recognition**]: Clustering – *algorithms*.

General Terms
Algorithms, Security.

Keywords
P-Sensitive *K*-Anonymity, Anonymization, User Constraints.

1. INTRODUCTION
Large amounts of personal data are constantly being collected in various application fields: healthcare, social networks etc. Besides its primary purpose for which it is collected in the first place, data is subsequently analyzed and mined. 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 against disclosure of confidential information [3].

A frequently used solution to protect microdata (a dataset where each tuple corresponds to one individual) is to modify it, in order to enforce an anonymity model on it. Some of the well-known anonymity models are *k*-anonymity, *l*-diversity, *p*-sensitive *k*-anonymity, *t*-closeness, etc.

We introduce in this paper the *constrained p-sensitive k-anonymity model*, that protects against identity and attribute disclosure [4], while keeping the quasi-identifiers' generalization restricted to certain user-specified boundaries. This model is an extension of constrained *k*-anonymity [5] and *p*-sensitive *k*-anonymity [6].

SAC'10, March 22-26, 2010, Sierre, Switzerland.

2. ANONYMITY AND CONSTRAINTS
Let $\mathcal{IM}$ be a microdata with the schema $\mathcal{R} = \{I_1, I_2,..., I_m, Q_1, Q_2,\ldots, Q_t, S_1, S_2,\ldots, S_r\}$ (identifiers, quasi-identifiers, and respectively sensitive attributes). Masked microdata $\mathcal{MM}$ (the microdata that will conform to the anonymity model) will have the schema $\mathcal{R}' = \{Q_1, Q_2,\ldots, Q_t, S_1, S_2,\ldots, S_r\}$, and at most $|\mathcal{IM}|$ tuples.

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 $\mathcal{MM}$.

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 μ, $MAGVal(\mu)$, is the value (leaf or not-leaf) in $\mathcal{H}_Q$ situated on the path from μ to the root, such that for any released microdata, the value μ is permitted to be generalized only up to $MAGVal(\mu)$ (see [5] for an example).

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

Definition 6 (*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.

3. ALGORITHM
Our approach to constrained *p*-sensitive *k*-anonymization partially follows an idea found in [1] and [2], 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 [5]). Our anonymization algorithm uses the concept of $\mathcal{MAM}$ defined and characterized in the followings.

Definition 7 (*Maximum Allowed Microdata*): The ***maximum allowed microdata*** for a microdata $\mathcal{IM}$, $\mathcal{MAM}$, is the masked microdata where every quasi-identifier value, μ, in $\mathcal{IM}$ is generalized to $MAGVal(\mu)$.

Property 1: For a given $\mathcal{IM}$, if its maximum allowed microdata $\mathcal{MAM}$ is not *p*-sensitive *k*-anonymous, then any masked microdata obtained from $\mathcal{IM}$ by applying generalization only will not satisfy constrained *p*-sensitive *k*-anonymity.

Property 2: If $\mathcal{MAM}$ satisfies *p*-sensitive *k*-anonymity then $\mathcal{MAM}$ satisfies the constrained *p*-sensitive *k*-anonymity property.

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 stated in Property 4. We will use the notation $\mathcal{OUT}$ to represent these entities.

Property 3: $\mathcal{IM} - \mathcal{OUT}$ can be masked using generalization only to comply with constrained *p*-sensitive *k*-anonymity.

Property 4: Any subset of $\mathcal{IM}$ that contains one or more entities from $\mathcal{OUT}$ cannot be masked using generalization only to achieve constrained *p*-sensitive *k*-anonymity.

Properties 3 and 4 show that $\mathcal{OUT}$ is the minimal tuple set that must be suppressed from $\mathcal{IM}$ 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 $\mathcal{OUT}$ set. Next, we create all *QI*-clusters in the maximum allowed microdata for $\mathcal{IM} - \mathcal{OUT}$. Last, each such cluster will be divided further, if possible, 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 [1] and a diversity measure [2].

The two stage constrained *p*-sensitive *k*-anonymization algorithm called *GreedyCPKA* is depicted in Figure 1. The numbers correspond to different values of a sensitive attribute – we only consider one sensitive attribute to make the process explainable graphically. The different geometrical shapes indicate tuples that belong to the same $\mathcal{MAM}$ cluster.

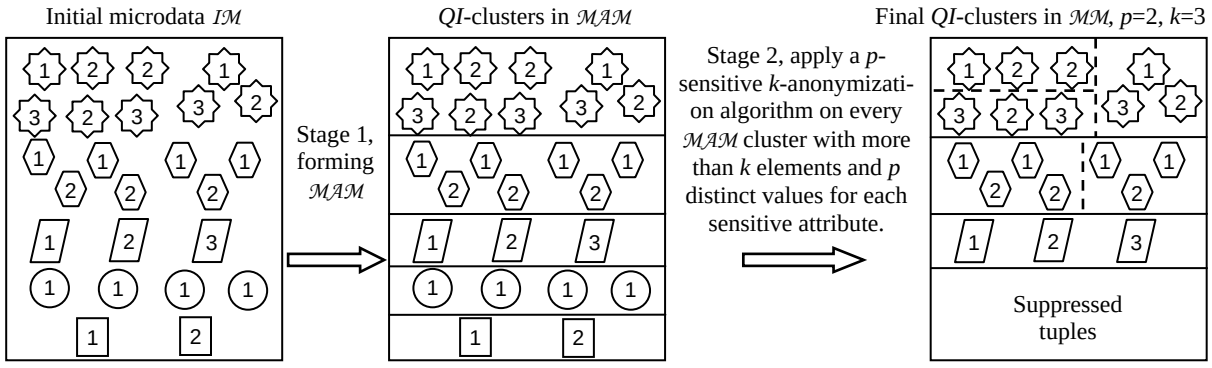

Figure 1. The two-stage process in creating constrained *p*-sensitive *k*-anonymous masked microdata.

4. RESULTS AND CONCLUSIONS

We compared the *GreedyCPKA* (generates constrained *p*-sensitive *k*-anonymous masked microdata), *GreedyCKA* (generates constrained *k*-anonymous masked microdata) [5], and *GreedyPKClustering* (generates unconstrained *p*-sensitive *k*-anonymous masked microdata) [2] 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 suppression amount performed by *GreedyCPKA* in order to produce constrained *p*-sensitive *k*-anonymous microdata in the presence of constraints.

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.

Acknowledgements: This work was partially supported by a NKU CINSAM grant.

5. REFERENCES

[1] Byun, J.W., Kamra, A., Bertino, E., and Li, N. Efficient *k*-Anonymization using Clustering Techniques. In *Proceedings of the DASFAA*, 188–200, 2006.

[2] Campan, A., Truta, T. M., Miller, J., and Sinca, R. A Clustering Approach for Achieving Data Privacy. In *Proceedings of the DMIN*, 321–327, 2007.

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

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

[5] Miller, J., Campan, A., and Truta, T. M. Constrained *K*-Anonymity: Privacy with Generalization Boundaries. In *Proceedings of the Workshop on Practical Preserving Data Mining with SIAM Statistical Data Mining*, 2008.

[6] Truta, T. M. and Bindu, V. Privacy Protection: *P*-Sensitive *K*-Anonymity Property. In *Proceedings of the Workshop on Privacy Data Management, with ICDE*, 94, 2006.