# Revisiting "Privacy Preserving Clustering by Data Transformation"

Stanley R. M. Oliveira[1], Osmar R. Zaïane[2]

[1] Embrapa Informtica Agropecuria, Caixa Postal 6041 - 13083-886 - Campinas, SP, Brazil
stanley@cnptia.embrapa.br
[2] Department of Computing Science, University of Alberta, Edmonton, Alberta T6G 2E8, Canada
zaiane@cs.ualberta.ca

**Abstract.** Preserving the privacy of individuals when data are shared for clustering is a complex problem. The challenge is how to protect the underlying data values subjected to clustering without jeopardizing the similarity between objects under analysis. In this short paper, we revisit a family of geometric data transformation methods (GDTMs) that distort numerical attributes by translations, scalings, rotations, or even by the combination of these geometric transformations. Such a method was designed to address privacy-preserving clustering, in scenarios where data owners must not only meet privacy requirements but also guarantee valid clustering results. We offer a detailed, comprehensive and up-to-date picture of methods for privacy-preserving clustering by data transformation.

Categories and Subject Descriptors: Information Systems [**Miscellaneous**]: Databases

Keywords: Privacy-preserving clustering, Hybrid Data Perturbation Method, Random projection, PPC solutions taxonomy.

## 1. INTRODUCTION

Understanding privacy in data mining requires understanding how privacy can be violated and the possible means for preventing privacy violation. In general, one major factor contributes to privacy violation in data mining: the misuse of data [Oliveira and Zaïane 2004c]. Users' privacy can be violated in different ways and with different intentions. Although data mining can be extremely valuable in many applications (e.g., business, medical analysis, etc), it can also, in the absence of adequate safeguards, violate informational privacy. Privacy can be violated if personal data are used for other purposes subsequent to the original transaction between an individual and an organization when the information was collected [Culnan 1993].

### 1.1 Defining Privacy for Data Mining

In general, privacy preservation occurs in two major dimensions: users personal information and information concerning their collective activity. We refer to the former as individual privacy preservation and the latter as collective privacy preservation, which is related to corporate privacy in [Clifton et al. 2002].

*Individual privacy preservation.* The primary goal of data privacy is the protection of personally identifiable information. In general, information is considered personally identifiable if it can be linked, directly or indirectly, to an individual person. Thus, when personal data are subjected to mining, the attribute values associated with individuals are private and must be protected from disclosure.

Miners are then able to learn from global models rather than from the characteristics of a particular individual.

*Collective privacy preservation.* Protecting personal data may not be enough. Sometimes, we may need to protect against learning sensitive knowledge representing the activities of a group. We refer to the protection of sensitive knowledge as collective privacy preservation. The goal here is quite similar to that one for statistical databases, in which security control mechanisms provide aggregate information about groups (population) and, at the same time, prevent disclosure of confidential information about individuals. However, unlike as is the case for statistical databases, another objective of collective privacy preservation is to protect sensitive knowledge that can provide competitive advantage in the business world.

### 1.2 Our Contribution in the Original Paper

We introduced a family of geometric data transformation methods (GDTMs) that distort confidential numerical attributes in order to meet privacy protection in clustering analysis. We benefited from the work on image processing [Gonzalez and Woods 1992]. Of particular interest is work on geometric transformation of digital images, notably the idea behind translation, scaling, and rotation. We also benefited from the work on statistical databases, particularly the intuition behind data distortion [Castano et al. 1995]. We showed that our transformation data methods are simple, independent of clustering algorithms, preserve the general features of the clusters, and have a sound mathematical foundation. Although our approach did not provide a comprehensive solution to the problem of privacy preservation in data mining, we argued that our approach was a simple building block toward privacy preserving data clustering. To date, such schemata had not been explored in detail in the context of Privacy Preserving Clustering (PPC).

Our proposed methods distort only confidential numerical attributes to meet privacy requirements, while preserving general features for clustering analysis. To our best knowledge this was the first effort toward a building block solution for the problem of privacy-preserving data clustering. The other approaches in the literature were restricted basically to address the privacy problem in the context of classification and association rules.

### 1.3 The Focus of This Short Paper

The work presented herein puts forward the need for new concepts and methods to address privacy protection against data mining techniques, notably in te context of data clustering.

We addressed a scenario in which some numerical confidential attributes of a database are distorted and made available for clustering analysis. In this context, users are free to use their own tools so that the restriction for privacy has to be applied before the mining phase on the data itself by data transformation.

The transformed database is available for secondary use and must hold the following restrictions: (1) the distorted database must preserve the main features of the clusters mined from the original database; (2) an appropriate balance between clustering accuracy and privacy must be guaranteed.

This paper updates, extends and reviews other PPC solutions by using data distortion, as a continuation of the original paper.

## 2. OUR CONTRIBUTION ON PPC

We now introduce a taxonomy including our methods for PPC by using data modification. The taxonomy basically encompasses three major categories: *Attribute Value Masking*, *Pairwise Object Similarity*, and *Attribute Reduction*.

*Attribute Value Masking.* This data transformation makes the original attribute values difficult to perceive or understand and preserves all the information for clustering analysis. Our data transformation that falls into this category is called Rotation-Based Transformation (RBT) [Oliveira and Zaïane 2004a]. The idea behind this technique is that the attributes of a database are split into pairwise attributes selected randomly. One attribute can be selected and rotated more than once, and the angle $\theta$ between an attribute pair is also selected randomly. RBT can be seen as a technique on the border with obfuscation. Obfuscation techniques aim at making information highly illegible without actually changing its inner meaning [Collberg et al. 1997]. In other words, using RBT the original data are masked so that the transformed data capture all the information for clustering analysis while protecting the underlying data values. One interesting application of RBT is privacy preservation of health data [Armstrong et al. 1999].

*Pairwise Object Similarity.* This technique is a data matrix representation in which a data owner shares the distance of the data objects instead of the location of the data points. This technique relies on the idea of the similarity between objects, i.e., a data owner shares some data for clustering analysis by simply computing the dissimilarity matrix (matrix of distances) between the objects and then sharing such a matrix with a third party [Oliveira and Zaïane 2004b]. This solution is simple and addresses PPC over centralized data. One of the most important advantages of this solution is that it can be applied to either categorical, binary, numerical attributes, or even a combination of these attributes. On the other hand, this solution can sometimes be restrictive since it requires a high communication cost.

*Attribute Reduction.* In this approach, the attributes of a database are reduced to a smaller number. The small number of attributes is not a subset of the original attributes since the transformation disguises the original attribute values by projecting them onto a random space. Our data transformation that lies in this category is called Dimensionality Reduction-Based Transformation (DRBT) [Oliveira and Zaïane 2007a; 2009]. This data transformation can be applied to both PPC over centralized data and PPC over vertically partitioned data. The idea behind this data transformation is that by reducing the dimensionality of a database to a sufficiently small value, one can find a trade-off between privacy and accuracy. Once the dimensionality of a database is reduced, the released database preserves (or slightly modifies) the distances between data points. In tandem with the benefit of preserving the similarity between points, this solution protects individuals' privacy since the underlying data values of the objects subjected to clustering are completely different from the original ones.

We first published the above taxonomy in [Oliveira and Zaïane 2007b]. We also published a complete taxonomy of privacy-preserving data mining methods in [Oliveira and Zaïane 2006], including methods for privacy preservation in classification, association rules, and data clustering.

## 3. CONCLUSIONS

This paper revisits privacy-preserving clustering by data transformation. We offered a detailed, comprehensive and up-to-date picture of methods for privacy-preserving clustering by data transformation.

The work presented herein puts forward the need for new concepts and methods to address privacy protection against data mining techniques, notably in data clustering. This paper updates, extends and reviews other PPC solutions by using data distortion, as a continuation of the original paper.

We hope that our study will be a useful reference for researchers and practitioners interested in the privacy aspects of data mining, notably those who work on privacy-preserving clustering (PPC).

## REFERENCES

Armstrong, M. P., Rushton, G., and Zimmerman, D. L. Geographically Masking Health Data to Preserve Confidentiality. *Statistics in Medicine* vol. 18, pp. 497–525, 1999.

Castano, S., Fugini, M., Martella, G., and Samarati, P. *Database Security*. Addison-Wesley Longman Limited, England, 1995.

Clifton, C., Kantarcioğlu, M., and Vaidya, J. Defining Privacy For Data Mining. In *Proceedings of the National Science Foundation Workshop on Next Generation Data Mining*. Baltimore, MD, USA, pp. 126–133, 2002.

Collberg, C., Thomborson, C., and Low, D. A Taxonomy of Obfuscating Transformations. Tech. rep., TR–148, Department of Computer Science, University of Auckland, New Zealand. July, 1997.

Culnan, M. J. How Did They Get My Name?: An Exploratory Investigation of Consumer Attitudes Toward Secondary Information. *MIS Quartely* 17 (3): 341–363, September, 1993.

Gonzalez, R. C. and Woods, R. E. *Digital Image Processing*. Addison-Wesley Publishing Company, 1992.

Oliveira, S. R. M. and Zaïane, O. R. Achieving Privacy Preservation When Sharing Data For Clustering. In *Proceedings of the Workshop on Secure Data Management in a Connected World, in conjunction with VLDB'2004*. Toronto, Ontario, Canada, pp. 67–82, 2004a.

Oliveira, S. R. M. and Zaïane, O. R. Privacy-Preserving Clustering by Object Similarity-Based Representation and Dimensionality Reduction Transformation. In *Proceedings of the Workshop on Privacy and Security Aspects of Data Mining, in conjunction with the Fourth IEEE International Conference on Data Mining*. Brighton, UK, pp. 21–30, 2004b.

Oliveira, S. R. M. and Zaïane, O. R. Toward Standardization in Privacy-Preserving Data Mining. In *Proceedings of the 3nd Workshop on Data Mining Standards, in conjuction with KDD 2004*. Seattle, WA, USA, pp. 7–17, 2004c.

Oliveira, S. R. M. and Zaïane, O. R. Privacy-Preserving Data Mining on the Web: Foundations and Techniques. In *Elena Ferrari; Bhavani Thuraisingham (Eds.). Web and Information Security. IDEA Group Inc.* pp. 282–301, 2006.

Oliveira, S. R. M. and Zaïane, O. R. A privacy-preserving clustering approach toward secure and effective data analysis for business collaboration. *Computers & Security* 26 (1): 81–93, 2007a.

Oliveira, S. R. M. and Zaïane, O. R. Privacy-Preserving Clustering to Uphold Business Collaboration: A Dimensionality Reduction-Based Transformation Approach. *International Journal of Information Security and Privacy* 1 (2): 13–36, 2007b.

Oliveira, S. R. M. and Zaïane, O. R. A Dimentionality Reduction-Based Transformation to Support Business Collaboration. In *Hamid R. Nemati (Eds.). Techniques and Applications for Advanced Information Privacy and Security: Emerging Organizational, Ethical, and Human Issues. IGI Global*. pp. 79–102, 2009.