Abstract No: SE-07 **Systems Engineering**

An efficient data perturbation scheme for preserving privacy on a numerical database

***Iuon-Chang Lin[1,2], Cheng-Yi Tsai[3] and Li-Cheng Yang[1]**

[1]Department of Management Information Systems,
National Chung Hsing University
[2]Department of Photonics and Communication Engineering, Asia University
[3]Department of Computer Science and Information Engineering, Asia University
**iclin@nchu.edu.tw*

Abstract

The data retention within an organization may increase rapidly with time. In order to reduce cost of organization, they may choose a third-party storage provider. There is a leakage crisis when provider cannot be trusted. Another scenario is a dealer collects all transaction data and provides it to a data analysis company for marketing purpose. For these reasons and beyons, preserving privacy in database becomes an important issue. This paper concerns the prediction of disclosure risk in numerical database. It presents an efficient noise generation that relies on Huffman coding algorithm and builds a noise matrix that can add noise intuitively to original value. Moreover, we adopt clustering technique before generating noise. The result shows that the running time of noise generation of clustering scheme is faster than non-clustering scheme.

Keywords: Database privacy, Disclosure control, Huffman coding, Micro aggregation, Noise matrix

Introduction

The data within an organization may increase rapidly. Instead of building up a storage space by itself, they may send these data into the data analysis company for some marketing purposes. Hence, the data mining techniques play an important role in the Knowledge Discovery in Databases (KDD). However, a malicious data analysis company may record personal data when organization publishes statistical database of the company. If the company is not trusted, there is a leakage crisis. For these reasons, privacy research has become popular in recent years. Statistical Data Bases (SDBs) are used to produce result of statistical aggregates such as sum, average, max and min. The results of statistical aggregates do not reveal the content about of any single individual tuple. However, the user may ask many legal queries to infer confidential information from gaining database responses.

In recent years, enhancement of the security of statistical database has received much attention. The problem of security in classical statistical database involves three different roles (Traub et al., 1984): statistician: who interest is to gain aggregate data; data owner: who desires individual records are security; database administrator: who needs to satisfy both of above roles. The privacy challenges in statistical database can be classified into two aspects (Lu & Tsudik, 2011): for data owner, he should avoid data theft by hackers, data abuse by service provider, and should restrict user access right; for user, it should hide query content, and database does not reveal query detail. There are many approaches that have been proposed. Navarro & Torra (2012) have introduced four categories as follows: a) *Perturbative methods,* which modify the original data to reach a degree of privacy. They usually called noise, b) *Non-perturbative methods,* the technique masks the data without introducing error. Data is not distorted, c) *Cryptographic methods,* which use classical cryptography system,