WP. 8
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Tarragona, Spain, 26-28 October 2011)

Topic (ii): Software research and development

Data Shuffling for Protecting Confidential Data

Prepared by Rathindra Sarathy, Oklahoma State University, U.S.A. and
Krishnamurty Muralidhar, University of Kentucky, U.S.A.

Data shuffling for protecting confidential data

Rathindra Sarathy[*] and Krish Muralidhar[**]

[*] Oklahoma State University, Stillwater, OK 74078 USA (rathin.sarathy@okstate.edu)
[**] University of Kentucky, Lexington, KY 40506, USA (krishm@uky.edu)

Abstract: Organizations gather and store extensive quantities of data, including sensitive information regarding individuals and other entities. Apart from transactional and operational purposes, analysis of such data is important to improving products and services and is often shared with other organizations. However, privacy concerns are leading to increasing calls for preventing the disclosure of sensitive confidential information contained in the data. Data shuffling can effectively respond to these concerns and to advance the ability of organizations to protect the privacy and confidentiality of data, while retaining its usefulness for analysis.

1 Introduction

Organizations derive many benefits from gathering, analyzing, and disseminating data regarding customers, suppliers, and other entities. Simultaneously, these activities raise issues of privacy and confidentiality of sensitive information. Unrestricted analysis, dissemination, and sharing of sensitive data could lead to disclosure of confidential information, so organizations need analytically valid data that does not disclose confidential information. Until recently, this problem was important only for a few governmental agencies (such as the Census Bureau) that released specialized data sets for sophisticated users. Recently, however, the scope of the problem has expanded to cover practically all organizations.

Research in statistical disclosure limitation techniques has led to the development of several tools and techniques that enable disseminated data to be analyzed while protecting individual privacy and confidentiality. We shall refer to these broadly as data masking techniques. Most of these techniques were developed in the context of data dissemination by a governmental agency. One such technique that enables data to be analyzed while preserving a high level of confidentiality is data perturbation. Perturbation techniques rely on "perturbing" or changing the original values in a manner that preserves analytical usability without compromising confidentiality. Unfortunately, many users look unfavorably (or possibly suspiciously) on values that have been "modified". This is likely to be particularly true of a typical user in commercial organizations who may not have the statistical sophistication of users of governmental data. Thus, techniques are needed that will foster greater acceptance of masked data among the common user.

One data masking approach with the potential to satisfy this requirement is data swapping. As the name implies, data swapping exchanges values of confidential variables between records without modifying the original values of the confidential variables. Data swapping has the intuitive appeal and ease of explanation that is not available with other masking techniques, but compared with perturbation methods, existing data swapping methods for numerical variables have low data utility or high disclosure risk (or both).

Data shuffling (Muralidhar and Sarathy 2006) combines the benefits of both perturbation and data swapping. Consequently, data shuffling offers high data utility like data swapping and also offers high security similar to perturbed data. This approach allows organizations to disseminate and share data for analysis, with minimal disclosure risk.

2 A Description of Data Shuffling

Data shuffling is based on the perturbation method proposed by Sarathy et al. (2002) using the multivariate copula to model the joint density of a data set where the variables have arbitrary marginal distributions and specified dependence characteristics. Let **X** represent a set of M confidential variables and let **S** represent a set of L non-confidential variables. **X** is assumed to be numerical while **S** is comprised of both categorical and numerical variables. Let **Y** represent the masked values of **X**. Let **R** represent the rank order correlation matrix of {**X**, **S**}. Define variables $\mathbf{X}^*$ and $\mathbf{S}^*$ as follows:

$$\begin{aligned} x^*_{i,j} &= \Phi^{-1}(F_{X_j}(x_j)),\ j = 1,...,M; i = 1,...,N, \text{and} \\ s^*_{i,k} &= \Phi^{-1}(F_{S_k}(s_k)), k = 1,...,L; i = 1, ...,N. \end{aligned} \tag{1}$$

where F_{X_j} and F_{S_k} represent the cumulative distribution functions of variables X_j and S_k, respectively, and Φ^{-1} represents the inverse of the standard normal distribution. The joint density of $\mathbf{X}^*$ and $\mathbf{S}^*$ is described by a multivariate standard normal distribution with correlation matrix $\boldsymbol{\rho}$, and the relationship between **R** and $\boldsymbol{\rho}$ can be described by:

$$\boldsymbol{\rho}_{ij} = 2\sin\left(\frac{\pi \mathbf{r}_{ij}}{6}\right) \tag{2}$$

where r_{ij} are the elements **R**. Since $\mathbf{X}^*$ and $\mathbf{S}^*$ have a joint multivariate normal distribution, it is now possible to generate the perturbed values $\mathbf{Y}^*$ as:

$$\mathbf{y_i^*} = \mathbf{\Sigma_{X^*S^*}(\Sigma_{S^*S^*})^{-1}s_i^*} + \mathbf{e_i}, \quad (3)$$

where $\mathbf{e_i} \sim \text{MVN}(\mathbf{0}, \mathbf{\Sigma_{X^*X^*} - \Sigma_{XS}(\Sigma_{S^*S^*})^{-1}\Sigma_{S^*X^*}})$. The values of $\mathbf{Y}^*$ can then be re-transformed back to the original marginal distribution of $\mathbf{Y^P}$ by:

$$y_{i,j}^{P} = F_{Xj}^{-1}(\Phi(y_{i,j}^{*})), i = 1,..., M;\ j = 1,..., N. \quad (4)$$

The copula-based perturbation approach requires that for a given data set {**X**,**S**} with a rank order correlation matrix **R**, the marginal distributions of **X** and **S** are known, a random observation can be generated from the specified marginal distribution. This step is needed to *generate* a new value for the perturbed variable using equation (4). However, in the data shuffling approach, we do not need to *generate* a new value; we only need the *rank* of the perturbed value. Hence, this procedure can be simplified further by using the following transformation in place of one used in equation (1):

$$s_{i,k}^{*} = \Phi^{-1}\left(\frac{(i) - 0.5}{N}\right), k = 1,..., L;\ j = 1, 2,..., N. \quad (5)$$

where (i) represents the rank order of $s_{i,k}$. The values of $\mathbf{y_i^*}$ are generated from:

$$\mathbf{y_i^*} = \mathbf{\rho_{X^*S^*}\rho_{S^*S^*}^{-1}(s_i^*)} + \mathbf{e_i} \quad (6)$$

where $e \sim \text{MVN}(\mathbf{0}, \mathbf{\rho_{X^*S^*}\rho_{S^*S^*}^{-1}\rho_{S^*X^*}})$. Since, rank of $y_{(i),j}^{*}$ = rank of $y_{(i),j}^{P}$, replacing $y_{(i),j}^{*}$ with $x_{(i),j}$, j = 1, …, M; i = 1,…, N results in **Y**. Note that equation (10) maintains conditional independence is maintained, ensuring that given **S**, **X** and **Y** are independent. Thus, the reverse-mapped values retain the same data utility and disclosure risk characteristics as the original perturbed values. If all the variables in the data set are confidential, an independent multivariate normal data set with correlation matrix **ρ** is generated and reverse mapping is performed on this data set. Simulation experiments to provide empirical verification of this result.

The data shuffling approach is non-parametric and does not require the identification of the marginal distributions of **X** and **S**, nor does it require *any* of the actual values in the data set to be used in the process of masking. The only data that is required is the rank order correlation matrix of the original variables and the ranks of the values of the non-confidential variables $s_{(i),j}$. In many cases, the owners of the data set may not have the necessary expertise to perform the shuffling. The non-parametric data shuffling approach provides a methodology by which the shuffling can be performed

securely using a third party that would never have access to *any* actual confidential data values.

3 Conclusions

In this paper, we have described a new shuffling procedure for masking confidential data. The advantages of this approach can be summarized as follows:

(1) The released data consists of the original values of the confidential variables (i.e., the marginal distribution is maintained exactly),
(2) All pair-wise monotonic relationships among the variables in the released data are the same as those in the original data, and
(3) Providing access to the masked microdata does not increase the risk of disclosure.

We will provide an extensive demonstration of the Data shuffling software.

References

Sarathy R., K. Muralidhar, R. Parsa. 2002. Perturbing non-normal confidential variables: The copula approach. *Management Science* **48** 1613-1627.

Muralidhar, K. and R. Sarathy. 2006. Data Shuffling - A New Masking Approach for Numerical Data. *Management Science* **52** 658-670.