Protecting Numerical Confidential Data using Data Shuffling: A Demonstration of Effectiveness of Approach and Flexibility of Delivery

Rathindra Sarathy
Ardmore Professor, Spears School of Business, Oklahoma State University, Stillwater OK 74078
rathin.sarathy@okstate.edu

Krish Muralidhar
Gatton Research Professor, Gatton College of Business & Economics, University of Kentucky, Lexington KY 40506
krishm@uky.edu

Abstract

Protecting numerical confidential data in data released by government agencies and other organizations poses a considerable challenge. There have been many different procedures that have been developed for this purpose. Among these procedures, Data shuffling offers the best compromise between preventing disclosure of information while preserving the usefulness of the data. Specifically Data shuffling offers the following advantages: (1) Individual data values are not modified but are shuffled between records thereby maintaining the distribution of individual variables exactly, (2) All linear and monotonic non-linear relationships between all variables are maintained providing a high level of analytical value, and (3) Providing the highest possible level of security (lowest level of both identity and value disclosure risk). In this demonstration, we provide an extensive analysis of the analytical value and disclosure risk characteristics of Data shuffling. We also demonstrate three modes of delivery based on the specific needs of the organization: (1) Excel based solution for small applications; (2) Web based solution for larger applications where the organizations wish to perform Data shuffling by themselves, and (3) Third party solution for organizations and agencies for large complex data sets. We believe that this demonstration will be useful for any organization that intends to analyze, share, or disseminate numerical confidential data without risk of disclosure.

Introduction

Organizations of all types can gather, store, and efficiently process large quantities of data. The primary purpose for gathering such data is to gain information from the data to improve business processes using statistical and data mining tools. The tremendous benefits of data mining have been repeatedly demonstrated in fraud detection, market-basket analysis, consumer profiling, anti-terrorism efforts, medicine, and many other domains. Well before data mining became popular, commercial organizations and government agencies were using statistical methods to analyze data to benefit consumers and society. Today, data mining draws from and adds to the many statistical analysis techniques. One of the key objectives of data mining is the discovery of new and useful relationships and patterns in the data. Some of these discoveries occur when data is mined specifically for the purposes of discovering such relationships. These unplanned discoveries are facilitated when users are provided access to the stored data.

Unfortunately, privacy and confidentiality issues are increasingly creating strong barriers that prevent us from realizing the full benefits of data. In many instances the data that was collected explicitly for analytical purposes sits in a secure facility where only a few authorized individuals are provided access to the data. Obviously this limits the usefulness of the data and defeats the very purpose for which they were gathered. Numerical data are of particular importance in this regard. They pose the greatest threat yet offer the greater benefits. They pose the greatest threat since they tend to be almost unique and an intruder with numerical data can easily compromise the privacy and confidentiality of sensitive records. They offer the greatest benefit since much of the business intelligence comes from numerical data. Hence, it is important

to protect numerical data from disclosure while also making it available for analysis purposes. This poses a serious dilemma in many organizational situations.

Data Shuffling

Data shuffling (US Patent# 7200757) developed by Muralidhar and Sarathy [2]offers an excellent solution to this dilemma. Data shuffling is a masking procedure where sensitive numerical attributes are shuffled between records. Based on advanced statistical modeling, Data shuffling is performed in such a manner so as to provide the highest possible level of protection from disclosure of sensitive information while simultaneously preserving the analytical value of the data by maintaining most of the relationships between the attributes. Specifically, Data shuffling maintains all linear and monotonic non-linear relationships among the masked variables to be the same as that between the original variables. This allows the user to analyze the data with the assurance that, for most statistical techniques, analyzing the masked data will provide very similar results as analyzing the original data. Among masking techniques for numerical data, data shuffling offers the highest level of protection while also providing the highest analytical value. In short, Data shuffling thwarts the attempts of any individual intent on compromising the data while rewarding the attempts of any individual interested in performing legitimate analysis.

Data Shuffling can be briefly described as follows. Consider a data set consisting of a set of (numerical) confidential variables X and a set of (numerical and categorical) non-confidential variables S. Data shuffling is implemented as follows.

(1) The rank order correlation of the entire data set is computed.
(2) Using the multivariate normal copula, the variables X and S are transformed to X* and S* such that they have a joint multivariate normal distribution. The product moment correlation of the new variables is computed using the rank order correlation computed in the first step [4].
(3) The perturbed values Y* are then computed using the general additive data perturbation method [5, 6].
(4) Let y*ij represent the perturbed value for the ith record and jth variable. In the original data set, replace y*(i),j with x(i),j ((i) = 1, 2, …, n; j = 1, 2, …,m) where y*(i),j and x(i),j represent the rank ordered observations of Y* and X, respectively.

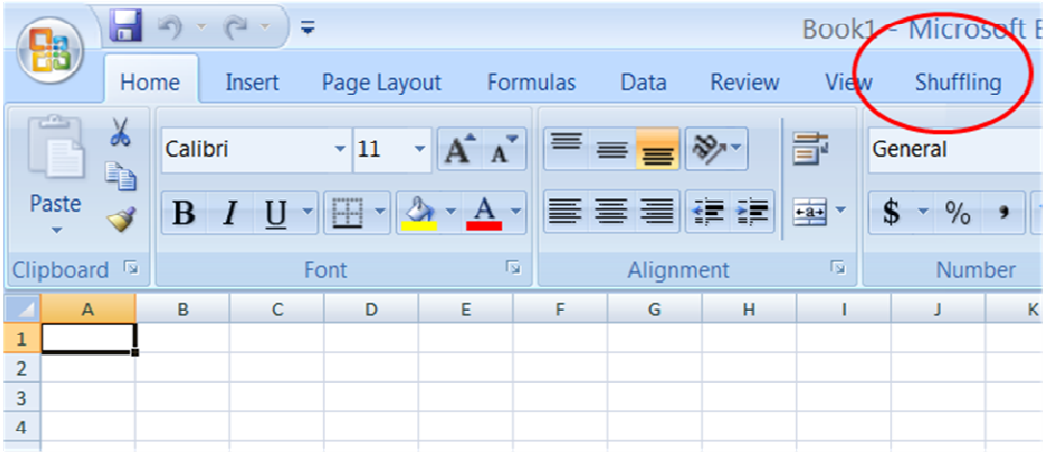
Muralidhar and Sarathy [1, 2] provide a comprehensive theoretical and empirical evaluation of the data utility and security of the data shuffling approach. The key characteristics of data shuffling can be summarized as follows. The shuffled values are actually the original values of the confidential variables assigned to a different observation. Hence, the univariate marginal distribution of the masked data is identical to that of the original data. The use of the copula-based perturbation approach enables data shuffling to maintain the rank order correlation of the masked data to be the same as that of the original data. This implies that data shuffling results in minimal information loss in linear and monotonic non-linear relationships among variables. For a complete discussion of data shuffling the interested reader is referred to [1, 2].

In this paper we demonstrate three modes of delivery based on the specific needs of the organization: (1) Excel based solution for small applications, (2) Web based solution for larger applications where the organizations wish to perform Data shuffling by themselves, and (3) Third party solution (installable java-based application) for organizations and agencies for large complex data sets. We believe that these modes of delivery will be useful for any organization that intends to analyze, share, or disseminate numerical confidential data without risk of disclosure.

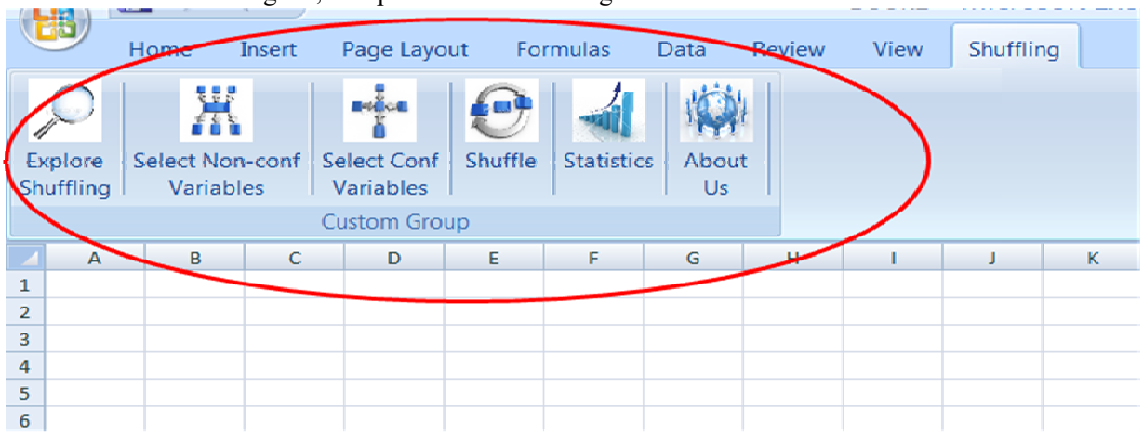
Modes of Delivery

Excel Add-in Based Solution

For small to medium data sets with up to 20 variables and 10,000 records, we have developed a Microsoft Excel Add-in. The Add-in is easy to install and implement. The installed Add-in option appears on the spreadsheet as follows:



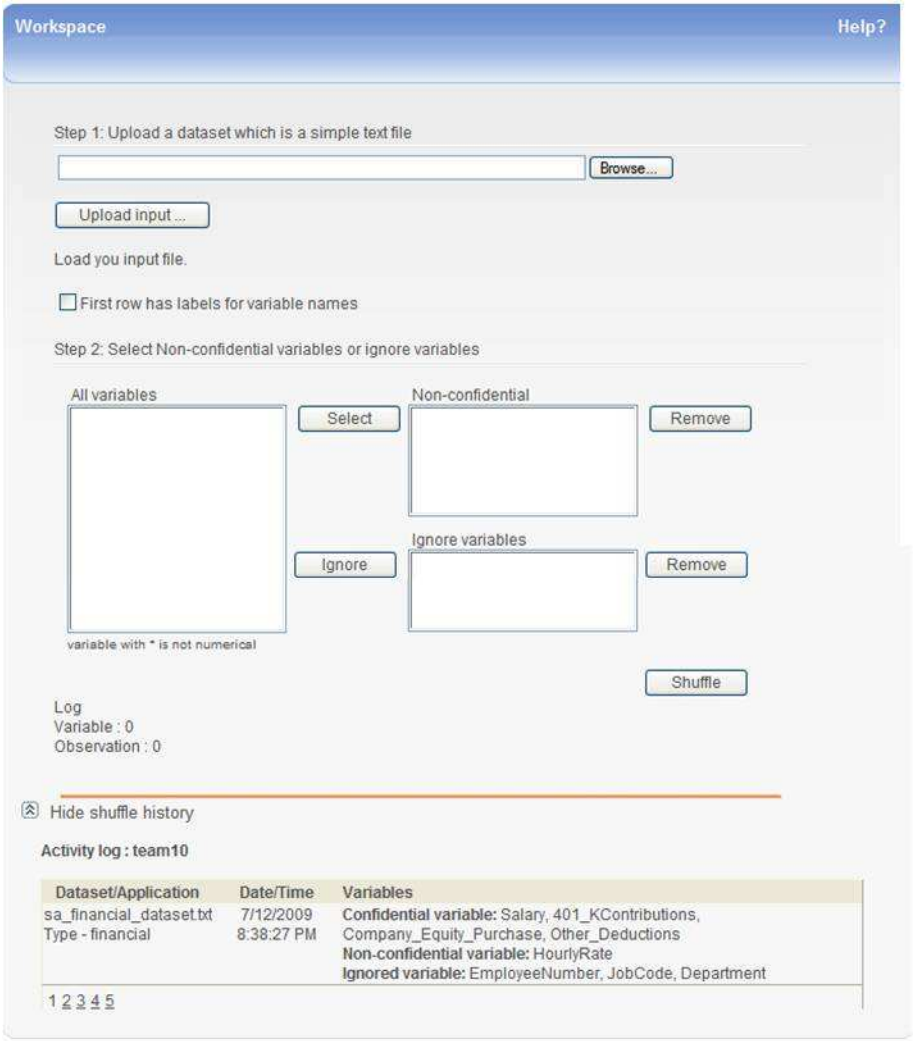
Selection of the Shuffling tab, will produce the following buttons:



As shown in the figure above, the user would be able to enter data into the spreadsheet, select the appropriate columns of non-confidential and confidential variables and perform shuffling. The statistics tab will produce basic statistics for both the original and the masked (shuffled) data to enable the user to assess the effectiveness of the shuffled data. Of course, an important advantage of the Excel Add-in is that the user may perform additional analysis using the Microsoft Excel software.

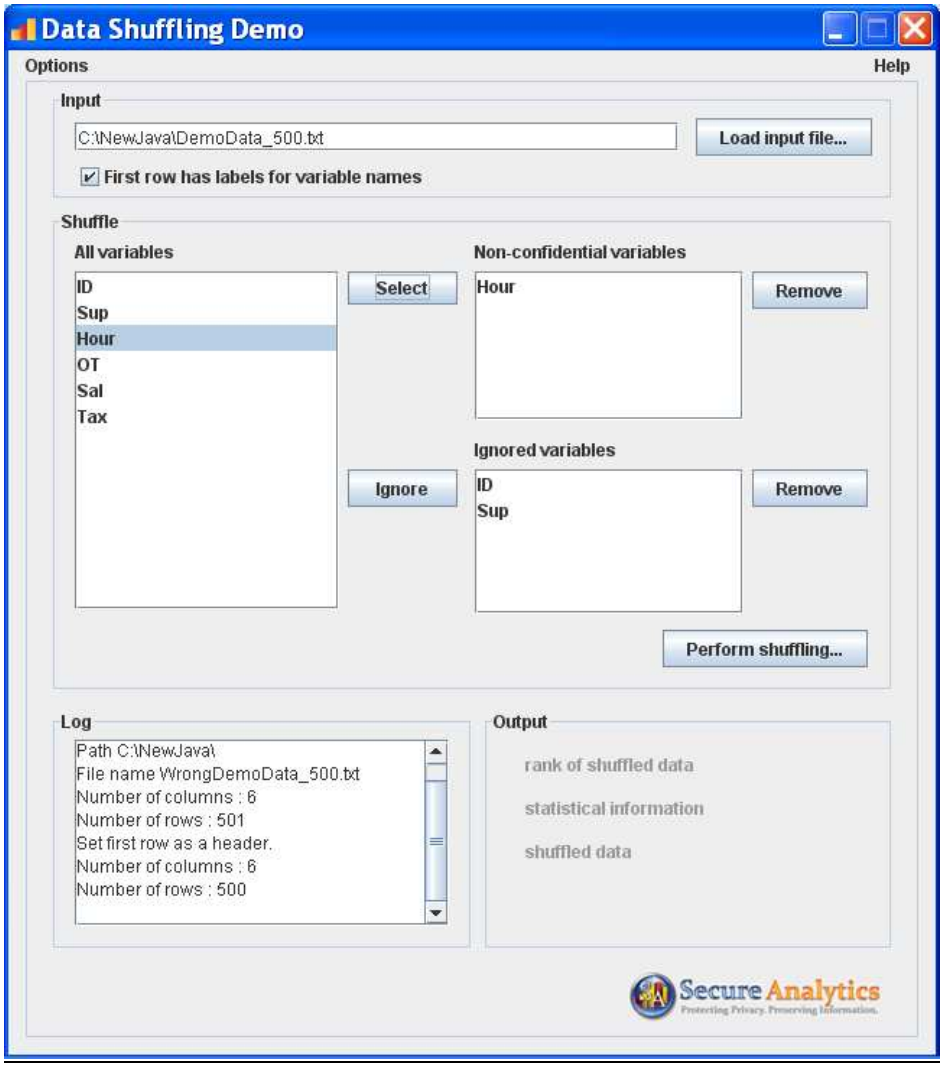
Web-Based Solution

Organizations that use medium-sized data sets (up to 100,000 records) and that only require masked data and do not want to install additional software, may take advantage of our web-based solution. The screen shots below show the web-based interface. As can be seen from the figure, the interface permits the uploading of a data set, the selection of appropriate variables to mask, and performing shuffling using a single button. One of the features of the web based solution is that it permits a user to store a history of their previous masking attempts, so that multiple runs (with different variables possibly masked each time) can be tracked easily.



Third-party solution (installable Java-based application)
The third mode of delivery that we visualize and demonstrate in the use of an installable application that is Java based. The java based application is similar to the web based solution except that it may be installed on the clients' site. Users who desire complete control over the data and the software will find this option useful. This mode also permits the use of third-party vendors who can provide both masking and analysis services to other organizations. The application can be built to scale for masking larger data sets.

The screen shot below shows the main screen of the application. It permits the choice of variables to ignore, variables that are non-confidential and variables that need to be masked. As with the web application, a single button click performs the data shuffling and saves both the masked output and comparative statistics from the original and the masked data.



Conclusions

In this paper we have briefly discussed a new data masking technique called Data Shuffling, the algorithm and three modes of delivery that organizations can employ related to data shuffling.

References

[1] K. Muralidhar and R. Sarathy, "A Theoretical Basis for Perturbation Methods," Statistics and Computing, vol. 13, pp. 329-335, 2003.

[2] K. Muralidhar and R. Sarathy, "Data shuffling - A new masking approach for numerical data," Management Science, vol. 52, pp. 658-670, 2006.

[3] L. T. Willenborg and T. D. Waal, Elements of statistical disclosure control. New York: Springer, 2001.

[4] R. Nelsen, "An introduction to Copulas," New York: Springer, 2007.

[5] K. Muralidhar, R. Parsa, and R. Sarathy, "A general additive data perturbation method for database security," Management Science, vol. 45, pp. 1399-1415, 1999.

[6] K. Muralidhar, R. Sarathy, and R. Parsa., "An improved security requirement for data perturbation with implications for e-commerce," Decision Sciences, vol. 32, pp. 683-698, 2001.