Performance Comparison of Privacy Preserving Perturbation algorithms in Association Rule Mining

[1]Vigneswari, [2]N.Komal Kumar, [3]G.V Bharath Kumar, [4]M. Vamsi Krishna
[1]Department of CSE,QIS Institute of Technology, Ongole, Andhra Pradesh, India
[2, 3, 4] Department of CSE,QIS Institute of Technology, Ongole, Andhra Pradesh, India
vigneswari121192@gmail.com, komalkumarnapa@gmail.com, bharath.gundapaneni9999@gmail.com,
vamsi.join@gmail.com

***Abstract*—This paper depicts the performance comparison of Non-Synthetic and Synthetic privacy preserving data perturbation algorithms. The perturbation algorithms are applied on different kinds of medical dataset which are then deployed on to the ARM(Association Rule Mining) and the experimental results are evaluated based on preserving privacy. The performance analysis is done by considering the algorithm which generates minimum lost rules and maximum Subterfuge rules that can be useful in preserving privacy.**

II.

***Index Terms*—Non-Synthetic, Synthetic, Privacy, Subterfuge.**

I. Introduction

Technical advances in privacy preserving algorithms makes the diplomatic and privileged information at risk. There are several mechanisms and techniques to provide solutions to privacy. An organization has many data sensitivity levels, misuse or modification of these sensitivity data can inimically affect the privacy of the organization and also international relations.A user can be given access to these data but that does not mean the user is a trusted entity, restricting the accessor shuffling [11] these data will not provide a solution for protecting a sensitive data.The main source of data inaccuracy is when the user provides wrong information. This is common when the user is asked with his/her personal information in public websites. For example, Pharmaceutical industry can ask the research vendors about some disease in order to calculate the correlation among the symptoms. When the confidential data are disclosed,affects the upcoming employment opportunities for the researches. But when the data values are masked, the industry might think that the data provided by the vendors are authentic that can be used for further process. Association rule hiding [10][12] involves in a process where some sensitive rules are suppressed. Richard Chow et.al[5] "Inference Problem" suggests that the user may infer sensitive data items for a non-sensitive ones. Regression analysis[4] techniques can be used to estimate the relationship among data items. Consider a case of information sensitivity in trade secrets of a Coca-Cola company, the formula for making Coca-Cola is not been widely spread and has proven effective tradeoff, even the secret behind the formula is not revealed under judges' order and considered to be a trade secret [8].

Suppose if the party wants to preserve the confidential information, first the information must be identified [13]. This can be accomplished by identifying the sensitive fields and performing perturbation technique on it.

II.PRIVACY PRESERVING PERTURBATION TECHNIQUES

In order to perform the masking of certain confidential values, a random noise [7] is used for perturbing the sensitive values of the original data items. Usually the masking deals with the numeric data values, since they are at most confidential at all cases; the data perturbation can be reversible or non- reversible unless certain parameters and methodologies are known. An efficient statistical method of privacy preserving perturbation for big data is also discussed in [3].

A. Synthetic data Perturbation

Synthetic data perturbation involves in addition or multiplying noise with the sensitive data values, which results in the immediate loss of information where the perturbation of the sensitive values occurs at different rates of mean and variance [6]. Synthetic data perturbation doesn't main the mean vector, so rather than addition of noise, multiplying will give an effective result which also provides better confidentiality.

1) Synthetic multiplicative perturbation:

Let P_{ij} be the value for the i^{th} Person with j^{th} characteristic, $i = 1, 2...N; j = 1, 2 ... N.$ and Noise $er_{i1}, er_{i2} \; . \; . \; . \; . \; er_{ip}$ corresponding to $x_{i1}, x_{i2} \; . \; . \; . \; , x_{ip}$.
er_{ij} is a random variable follows an uniform distribution with mean μ_j=0 and variance σ_j=1
Synthetic (P, er)
begin

```
Let $P_i=\{P_1,P_2,P_3...P_n)$ be sensitive values for all Person $(P_i)$
Compute $er_{ij}$ by uniform distribution with mean=0 and some variance
Calculate $z_{ij} = P_{ij}* er_{ij}$
return $z_{ij}$
end
```

2)Synthetic logarithmic transformation perturbation:

```
Synthetic Logarithmic Transformation (P,e)
begin
Let $P_i=\{P_1,P_2,P_3...P_n)$ be sensitive values for all Person $(P_i)$
Calculate the error 'er' by an uniform distribution (mean=0, variance=1)
Compute $y_{ij}$by taking logarithm for Person $i^{th}$ value $P_i$ and add the error 'er'
Calculate the antilog for $y_{ij}$
return $z_i$
end
```

B. Non- Synthetic data Perturbation

Non-synthetic data perturbation [1] deals with two most important parameters, such as a confidential parameter denoted by α and a single non confidential variable β. For simplicity we assume that mean between α and β equal 0. The parameter ξ is a "resemblance parameter". When the resemblance parameter is zero (i.e. ξ=0), the confidential α and the resultant z are most dissimilar, When the resemblance parameter increases to ξ=1, the confidential α and the resultant z are most similar. Thus the resemblance parameter allows the data user to control the sensitivity levels of the data values.

From Muralidhar [6], We begin to perturb the values using the formula

$$z_i= \xi\, \alpha_i+(1-\xi)\, \mu\beta_i+\text{sqrt}((1-\xi^2)(1-\mu^2))\rho \qquad (1)$$

```
Non synthetic(ξ, α, β, ρ, μ)
Begin
Let $\alpha_i=\{\alpha_1,\alpha_2,\alpha_3,....\alpha n\}$ be confidential values, $\beta_i=\{\beta_1,\beta_2,\beta_3,....\beta n\}$be non-confidential,
Compute $\rho$, normally distributed with mean 0 and unit variance.
Calculate μ, correlation between $\alpha$ and $\beta$
For all $\alpha_i$
    Compute $z_i$ with ξ
End
```

Consider a simple case of dataset containing some 20 observations for 'α' (Confidential), 'β' (Non-Confidential) and the similarity parameter values 'ξ'.correlation between α and β is μ and ρ is normally distributed with mean 0 and unit variance.

When ξ = 1, we get z = 0, the coefficient of ρ is also zero, resulting in $z_i = \alpha_i$, which is the equivalent of releasing the unmodified values of α.

When, when ξ = 0, the perturbed values z_i are not a function of the confidential value α_i, but is a function of only β and ρ implying that the values are modified

A simulated data set with these characteristics is presented in Table 1

TABLE 1
ORIGINAL AND PERTURBED DATA VALUES SHOWING VARIATIONS WITH RESPECT TO 'ξ'

	ξ	0.2	0.4	0.6	0.9
β	α	z	z	z	z
8.1	5.1	2.2	4.0	4.6	5.0
8.6	4.3	2.1	2.9	3.7	4.2
11	7.7	4.3	5.4	6.4	7.5
8.7	6.5	4.5	5.2	5.8	6.3
7.6	5.0	2.3	2.9	3.8	4.8
13.7	10.8	7.9	8.4	9.2	10.6
9.8	9.0	6.3	6.5	7.5	8.9
9.0	4.8	2.0	2.9	3.9	4.6
9.3	7.3	5.7	6.0	6.3	7.0

Consider a case where the resemblance parameter is varied from 0.9 to 0.3, the resemblance parameter starts from 0.9, the values of αi and zi are mostly same, the variation of the resemblance parameter is continued from 0.6 to 0.2, the variation in the confidential and the resultant values are recorded, the values of the resultant eventually comes closer when the resemblance parameter is high (0.9), and goes to an higher difference when the resemblance parameter is low (0.2). so the resemblance parameter can be used to fine tune the confidentiality of the parameters.

III. EXPERIMENTAL ANALYSIS

The proposed methodology is carried out with the help of Apriori association rule mining algorithm and performance of Non synthetic and Synthetic data perturbation algorithms are analyzed [2]. The datasets used in this experimental analysis are the birth and death measure from U.S Department of Health and Human Services and heart disease dataset from UCI repository. The birth and death dataset contained twenty one quantative and seven categorial attributes and heart disease dataset contained sixteen quantative and eight categorial values. For simplicity we used only seven quantative, three categorial and six quantative, 4 categorial values in analysing the performance of the perturbation algorithms in association rule mining.

The performance analysis of these algorithms are considered based on higher subterfuge rules and less production of lost rules capabilities. Higher the subterfuge rules makes the data more confidential by hiding the rules which ultimately provides the privacy for the data. Increase in the production of lost rules will lead to lesser data utility. Hence the lost rules should always be minimum.

Synthetic multiplicative, Synthetic logarithmic and non-synthetic data perturbation algorithms are tested in the analysis and the following results are obtained.

Non-Synthetic perturbation results are as follows:

Figure 1 shows the Number of rules generated for varying confidence when ξ=0.2 with support=10.

Figure 2 shows the Number of rules generated for varying confidence when ξ=0.4 with support=10.

Figure 3 shows the Number of rules generated for varying confidence when ξ=0.6 with support=10.

Figure 4 shows the Number of rules generated for varying confidence when ξ=0.9 with support=10. The scenarios of the figures stipulate that tuning of the resemblance parameter changes the privacy of data as well as rules. Lesser the ξ parameter higher the sensitivity and vice versa.

Figure 5 shows the No of rules generated for varying confidence and resemblance parameter 'ξ' for a constant support of 10.

Figure 6 shows the No of lost rules generated for varying confidence and resemblance parameter ξ=0.2 is considered for comparison since lesser the value of resemblance parameter higher the sensitivity for a constant support of 10.

Figure 7 shows the No of subterfuge rules generated for varying confidence and resemblance parameter ξ=0.2 is considered for comparison since lesser the value of resemblance parameter higher the sensitivity for a constant support of 10.

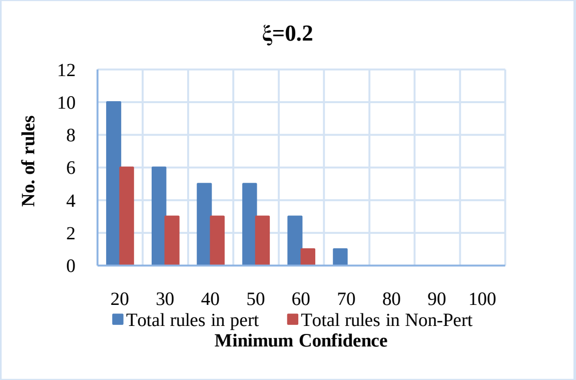

Figure 1. Varying confidence Vs. No of Rules

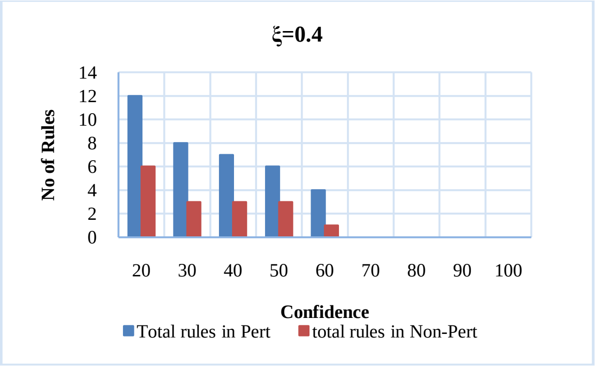

Figure 2. Varying confidence Vs. No of Rules

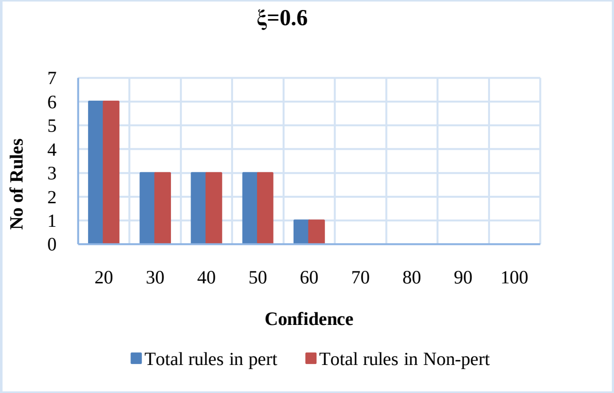

Figure 3. Varying confidence Vs. No of Rules

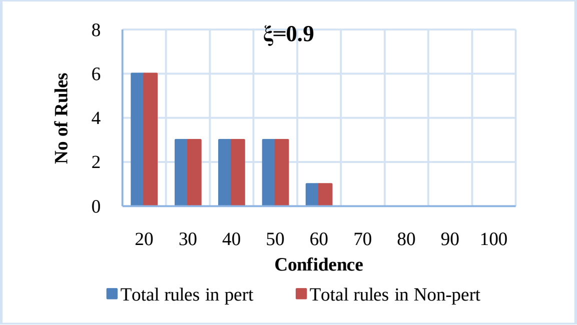

Figure 4. Varying confidence Vs. No of Rules

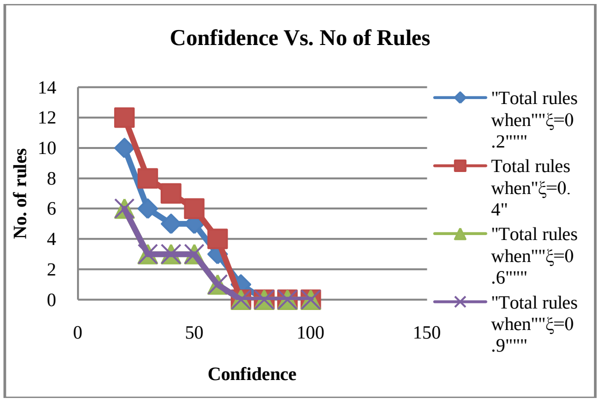

Figure 5. Varying confidence Vs. No of Rules with varying 'ξ'

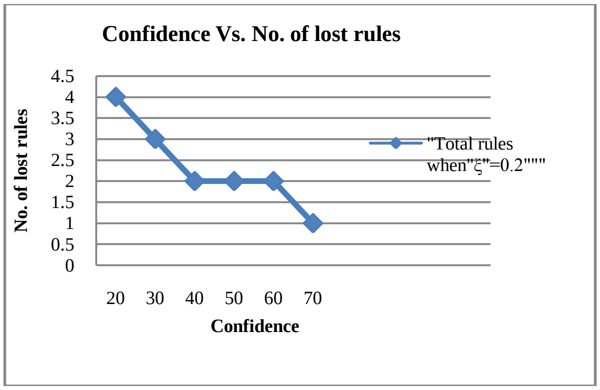

Figure 6. Varying confidence Vs. No of lost Rules with 'ξ=0.2'

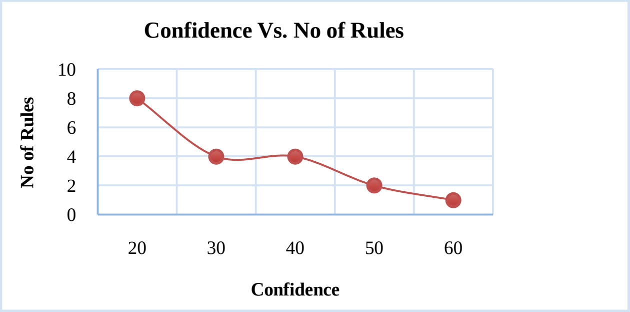

Figure 7. Varying confidence Vs. No of subterfuge rules with 'ξ=0.2'

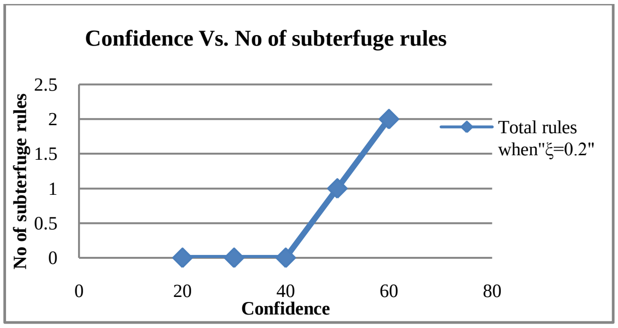

Figure 8. Varying confidence Vs. No of rules

Synthetic multiplicative perturbation results are as follows: Figure 8 shows the No of rules generated with varying confidence

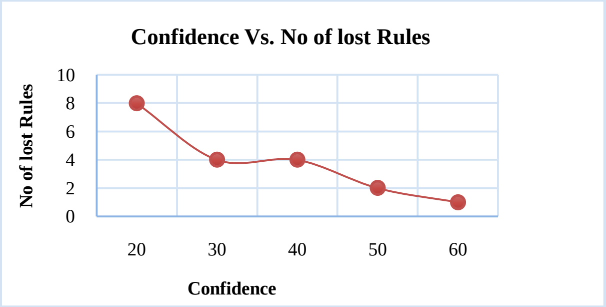

Figure 9. Varying confidence Vs. No of lost rules

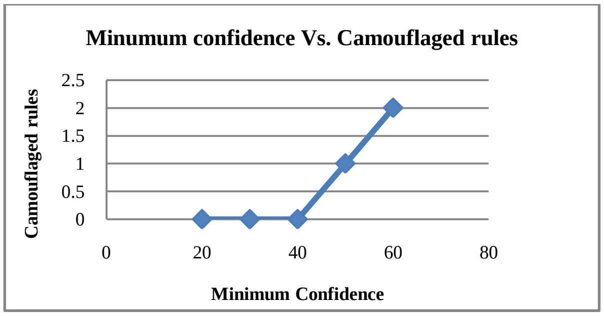

Figure 10. Varying confidence Vs. No of subterfuge rules

Figure 9 shows the No of lost rules generated with varying confidence. Figure 10 shows the No of subterfuge rules generated with varying confidence.

Synthetic logarithmic perturbation results are as follows:

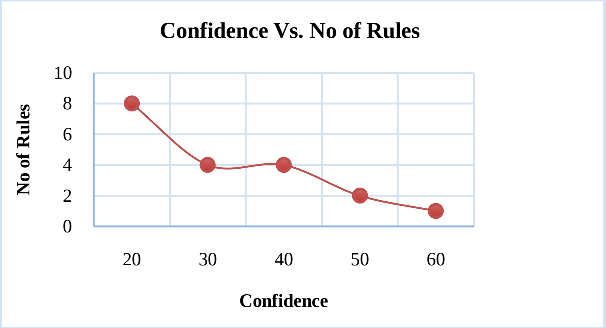

Figure 11. Varying confidence Vs. No of rules

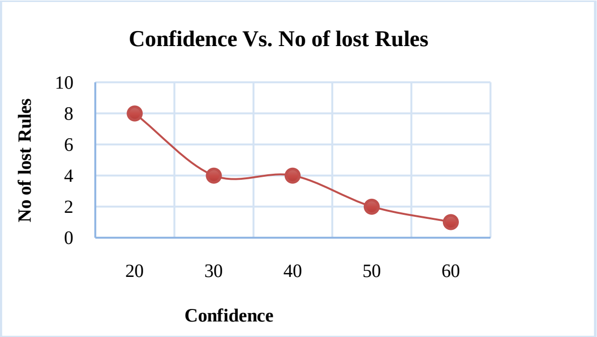

Figure 12. Varying confidence Vs. No of lost rules

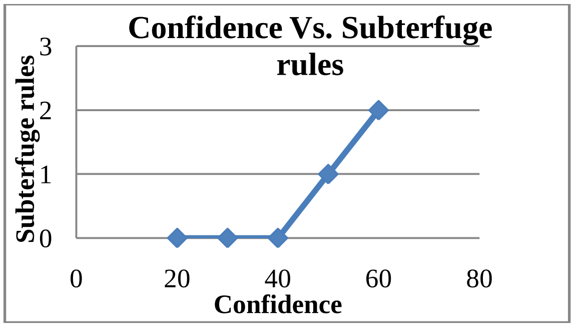

Figure 13. Varying confidence Vs. Noof subterfuge rules

Figure 11 shows the No of rules generated with varying confidence. Figure 12 shows the No of lost rules generated with varying confidence. Figure 13 shows the No of subterfuge rules generated with varying confidence.

TABLE 2
EVALUATION OF NON SYNTHETIC, SYNTHETIC MULTIPLICATIVE AND SYNTHETIC LOGARTHMIC ALGORITHMS ON ASSOCIATION RULES

Algorithm used	Confidence	No of Rules generated	No of lost rules	No of Subterfuge rules
Non Synthetic	20	12	4	2
	30	8	3	3
	40	7	2	3
	50	6	2	4
	60	4	2	5
Synthetic multiplicative	20	8	8	0
	30	4	4	0
	40	4	4	0
	50	2	2	1
	60	1	1	1
Synthetic logarthmic	20	8	8	0
	30	4	4	0
	40	4	4	0
	50	2	2	1
	60	1	1	2

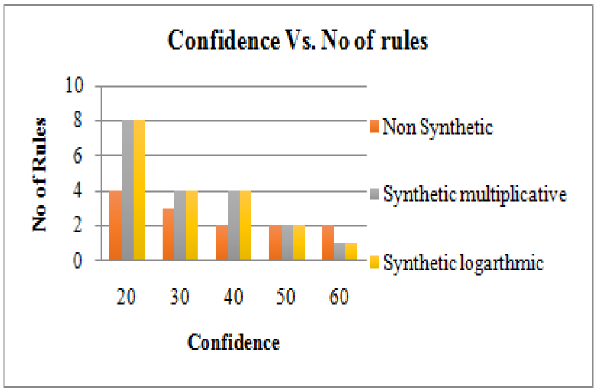

Figure 14. Varying confidence Vs. No of rules

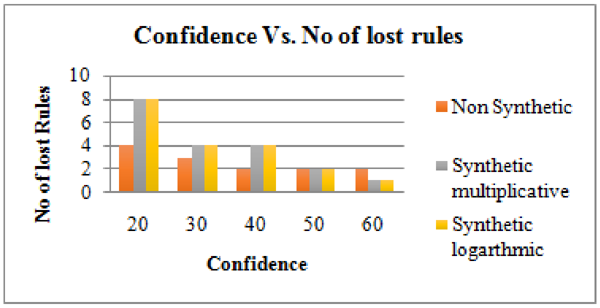

Figure 15. Varying confidence Vs. No of lost rules

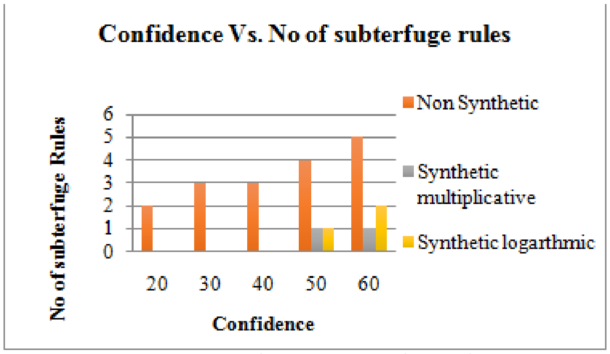

Figure 16. Varying confidence Vs. No of subterfuge rules

Figure 14 shows the varying confidence with No of rules generated for Non synthetic, synthetic multiplicative and synthetic logarithmic algorithms. The scenario depicts that both synthetic multiplicative and synthetic logarithmic algorithms produces excess rules by which the algorithm becomes inefficient.

Figure 15 shows the varying confidence with No of lost rules. The scenario depicts that lost rules are higher in case of synthetic multiplicative and synthetic logarithmic when compared to Non synthetic perturbation.

Figure 16 shows the varying confidence with subterfuge rules. The scenario depicts that non-synthetic data perturbation algorithm generates more subterfuge rules when compared to other two synthetic algorithms in analysis.

IV. CONCLUSION

The evaluation of these three algorithms shows that two synthetic data perturbation algorithms results are impotent, by producing many unfavorable rules, which makes the data to less utilizable. But Non-synthetic data perturbation algorithm produces efficacy rules which maintains the maximum utility and greater privacy [14]. The synthetic algorithms provides lesser subterfuge rules that leads to the inefficiency of the algorithm. Where as the non-synthetic algorithm is having higher subterfuge rules resulting in perpetuity of the data that imposes the privacy on it.

Finally by evaluating the results we have perceived that the non-synthetic perturbation algorithm has much efficiency than that of synthetic perturbation algorithms. The future work can be extended by deploying the non-synthetic perturbation algorithm on to the cloud environment.

REFERENCES

[1]Dr.T.Ravi, R. Prasanna Kumar, KomalKumar.N,"A Non Synthetic Data Perturbation Technique for Privacy preservation in Association Rule Mining", In the International Journal of Applied Engineering Research, 2014, 9(24): 24311-24320

[2] Dr. T. Ravi, R. Prasanna Kumar, Komal Kumar. N., G. Ragu."Synthetic Data Perturbation Techniques for Privacy Preservation in Association Rule Mining", In Journal of Applied Sciences and Research, 11(10), 2016, pg 55-59.

[3] P. Shobha Rani and D. Vigneswari, "An Efficient Statistical Method for Providing Privacy and Security in Big Data", In World Engineering & Applied Sciences Journal, 2016, 7 (4): 235-240.

[4] Armstrong, J. Scott ,"Illusions in Regression Analysis". International Journal of Forecasting (forthcoming), 2012, 28 (3): 689. doi:10.1016/j.ijforecast.2012.02.001

[5] Richard Chow,"Privacy Leaks Using Corpus-based Association Rules", KDD'08, Las Vegas, Nevada, USA, 2008, pp. 24-27.

[6] Muralidhar, K. and R. Sarathy, "Generating Sufficiency- based Non-Synthetic Perturbed Data," Management Science, Transaction on DataPrivacy, 2008, pp. 17-33.

[7] J.J. Kim and W.E. Winkler, "Multiplicative Noise for MaskingContinuous Data," Technical Report Statistics #2003-01, Statistical Research Division, US Bureau of the Census, Washington D.C.], 2003.

[8] For God, Country & Coca-Cola, by Mark Pendergrast, 2nd Ed., Basic Books 2000, p. 456.

[9] Evfimievski, A., R. Srikant, R. Agrawal and J. Gehrke, "Privacy Preserving Mining of Association Rules," Proc. Eighth ACMSIGKDD Int‟l Conf. Knowledge Discovery and Data Mining (KDD‟02). 2002.

[10] ArisGkoulalas–Divanis;Vassilios S. Verykios, "Association Rule Hiding For Data Mining" Springer, DOI 10.1007/978-1-4419-6569-1, Springer Science + Business Media, LLC, 2010.

[11] Muralidhar, K. and R. Sarathy, "Data Shuffling- A New Masking Approach for Numerical Data," Management Science, 2006, 52(5): 658- 670.

[12] Agrawal, R., T. Imieliński, A. Swami, 1993. "Mining association rules between sets of items in large databases". "Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93". pp: 207. Agrawal, R. and R. Srikant, 2000.

[13] S. Warner, "Privacy-Preserving Data Mining", SIGMOD, "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias," J. Am. Statistical Assoc., 60: 63-69., 1965, pp: 161-172.

[14] J. Vaidya and C. Clifton. Privacy preserving association rule mining invertically partitioned data. Proc. of 8th ACM Intl. Conference on Knowledge Discovery and Data Mining (KDD), 2002.