Privacy-Preserving Data Mining for Horizontally-Distributed Datasets using EGADP

Mohammad Saad Al-Ahmadi, King Fahd University of Petroleum and Minerals (KFUPM), Dhahran, Saudi Arabia, alahmadi@kfupm.edu.sa, dr.alahmadi@gmail.com

Abstract

In this paper, we investigate the possibility of using EGADP for protecting data in horizontally-distributed datasets. EGADP [10] is a new advanced data perturbation method that masks confidential numeric attributes in original datasets while reproducing all linear relationships in masked datasets. It is developed for centralized datasets that are owned by one owner, and no study (to the best of our knowledge) suggests and investigates empirically the possibilities of using it to protect distributed confidential datasets. This study is intended to fill this gap.

1. Introduction

Data mining techniques and algorithms play a central role in knowledge discovery. Data miners frequently need full access to the data in order to build accurate models. However, one of the biggest barriers facing data mining projects today is the "*inability to release data*" due to privacy concerns [3].

Additionally, the required data can be found with more than one party. In this situation there are two possibilities; datasets can be horizontally or vertically divided [5, 11]. In vertically-distributed datasets, different parties have access to all records but each party owns different attributes (or variables). In horizontally-distributed datasets, contributing parties own all required attributes but they have a subset of the needed records.

Motivation Example

Assume there are three banks that are participating in an alliance that allows them to share data about their customers. The shared data contains three non-confidential attributes **S** (S_1: categorical either 0 or 1 (e.g. male or female), and S_2 and S_3: numeric) and three confidential numeric attributes **X** (X_1, X_2 and X_3).

The banks want to share and combine their horizontally-distributed datasets into one integrated dataset. The goal of sharing and integrating their distributed data is to *independently* perform basic statistical analysis and data mining tasks, such as linear multiple regressions, without restriction for the mutual benefit of all banks. At the same time, the banks must protect the sensitive numeric attributes (i.e. **X**) before sharing and building the required integrated dataset.

For this motivation example, we simulated a new multivariate normally-distributed dataset similar to the one used in [8]. It consists of 6 attributes and 50,000 records. Then, we randomly divided the records of this dataset into three (sub-)datasets. Each dataset belongs to a different bank identified by 'Bank No' attribute. Table 1 lists few record examples from this simulated integrated dataset, which represents the combined three banks' datasets, along its assumed source bank. The contributions of Bank 1, Bank 2, and Bank 3 to this dataset are 18,000, 20,000 and 12,000 records, respectively.

Table 2 presents the statistical measures for the simulated 50,000-records dataset. These measures include mean, standard deviation, correlation matrix, covariance matrix, and number of records. As we will discuss later, these measures are very important for data utility when we mask data using EGADP. Additionally, Table 3, Table 4, and Table 5 present similar measures for Bank 1, Bank 2, and Bank 3 (sub-)datasets, respectively.

Table 1: Original Dataset Combining the Three Banks' Datasets

	Original Dataset*						
Customer	**Non-Confidential Attributes (S)**			**Confidential Attributes (X)**			**Bank**
No	S_1	S_2	S_3	X_1	X_2	X_3	**No**
1	1	102.810	38.468	65.156	21.467	50.683	2
2	1	59.328	36.755	56.266	17.991	46.043	1
3	0	101.840	54.821	87.120	21.335	52.449	3
4	1	99.733	51.979	74.428	23.664	46.659	1
5	1	70.282	53.541	77.308	7.459	47.608	2
:	:	:	:	:	:	:	:
49,996	0	109.810	53.500	113.900	12.272	39.900	1
49,997	0	134.430	61.147	90.463	28.679	67.198	2
49,998	1	101.970	57.772	91.724	19.943	54.317	3
49,999	1	93.589	47.022	68.249	12.492	44.614	2
50,000	1	99.582	53.304	88.220	17.811	48.313	1

* The unit of S_2, S_3, X_1, X_2, X_3 is \$000

Table 2: Statistical Measures for the Original Three Banks' Datasets (Combined)

Summary Statistics		Non-Confidential Attributes S			Confidential Attributes X		
		Correlation					
Mean	Attribute	S_1	S_2	S_3	X_1	X_2	X_3
0.50	S_1	1					
100.00	S_2	-0.00023	1				
50.00	S_3	-0.00214	0.700	1			
80.00	X_1	-0.00295	0.800	0.750	1		
20.00	X_2	0.00184	0.500	0.400	0.250	1	
50.00	X_3	0.00689	0.300	0.200	0.150	0.600	1
Standard Deviation	Attribute	Covariance					
		S_1	S_2	S_3	X_1	X_2	X_3
0.50	S_1	0.250					
20.00	S_2	-0.002	400.000				
10.00	S_3	-0.011	140.000	100.000			
20.00	X_1	-0.029	320.000	150.000	400.000		
5.00	X_2	0.005	50.000	20.000	25.000	25.000	
10.00	X_3	0.034	60.000	20.000	30.000	30.000	100.000

Dataset Size: 50,000 records

Table 3: Statistical Measures for Bank 1's Original Dataset

Summary Statistics		Non-Confidential Attributes S			Confidential Attributes X		
		Correlation					
Mean	Attribute	S_1	S_2	S_3	X_1	X_2	X_3
0.50	S_1	1					
100.17	S_2	-0.00650	1				
50.07	S_3	-0.00511	0.700	1			
80.18	X_1	-0.01097	0.798	0.748	1		
20.03	X_2	0.00094	0.507	0.402	0.252	1	
50.08	X_3	0.00292	0.304	0.204	0.151	0.604	1
Standard Deviation	Attribute	Covariance					
		S_1	S_2	S_3	X_1	X_2	X_3
0.50	S_1	0.250					
19.94	S_2	-0.065	397.720				
10.02	S_3	-0.026	139.780	100.320			
19.91	X_1	-0.109	317.040	149.270	396.570		
5.02	X_2	0.002	50.778	20.194	25.195	25.201	
10.01	X_3	0.015	60.599	20.419	30.087	30.362	100.120

Dataset Size: 18,000 records

Table 4: Statistical Measures for Bank 2's Original Dataset

Summary Statistics		Non-Confidential Attributes S			Confidential Attributes X		
		Correlation					
Mean	Attribute	S_1	S_2	S_3	X_1	X_2	X_3
0.50	S_1	1					
99.90	S_2	0.00405	1				
49.99	S_3	0.00100	0.699	1			
80.01	X_1	0.00212	0.802	0.752	1		
19.93	X_2	0.00143	0.494	0.396	0.246	1	
49.88	X_3	0.01529	0.295	0.190	0.143	0.598	1
Standard Deviation		**Covariance**					
	Attribute	S_1	S_2	S_3	X_1	X_2	X_3
0.50	S_1	0.250					
20.11	S_2	0.041	404.380				
10.03	S_3	0.005	141.030	100.660			
20.16	X_1	0.021	325.120	152.030	406.430		
4.99	X_2	0.004	49.572	19.828	24.784	24.949	
9.99	X_3	0.076	59.208	19.029	28.834	29.839	99.879

Dataset Size: 20,000 records

Table 5: Statistical Measures for Bank 3's Original Dataset

Summary Statistics		Non-Confidential Attributes S			Confidential Attributes X		
		Correlation					
Mean	Attribute	S_1	S_2	S_3	X_1	X_2	X_3
0.50	S_1	1					
99.92	S_2	0.00216	1				
49.92	S_3	-0.00278	0.702	1			
79.71	X_1	0.00075	0.799	0.750	1		
20.07	X_2	0.00400	0.500	0.405	0.254	1	
50.08	X_3	-0.00097	0.303	0.212	0.160	0.597	1
Standard Deviation		**Covariance**					
	Attribute	S_1	S_2	S_3	X_1	X_2	X_3
0.50	S_1	0.250					
19.90	S_2	0.022	396.110				
9.92	S_3	-0.014	138.610	98.429			
19.86	X_1	0.007	315.900	147.690	394.370		
4.98	X_2	0.010	49.541	20.002	25.087	24.773	
10.00	X_3	-0.005	60.396	20.990	31.823	29.707	100.000

Dataset Size: 12,000 records

2. Mask and Perturb Data using EGADP

GADP [8], IPSO [2], and EGADP [10] are advanced data perturbation methods that were developed mainly to mask sensitive *numeric* attributes while reproducing original *linear* relationships in masked datasets. In order to guarantee that all relationships are linear, original datasets are assumed to be multivariate normally distributed [6].

EGADP data perturbation method avoids the problems found in the other two advanced data perturbation methods. Muralidhar and Sarathy [10] proved that EGADP does not suffer from a sampling problem as in the case of GADP even when EGADP is used to mask very small datasets. They also showed that IPSO suffers from a security problem that EGADP avoids.

The procedure of masking datasets using EGADP can be explained as follows [10]:

1. Regress $\mathbf{X}$ on $\mathbf{S}$ to calculate the fitted values **u** and the residuals set **r**, which is independent of **S**.
2. Compute the covariance matrix $\sum_{\mathbf{r}}$ of the residuals set **r**. This covariance matrix will be used later to scale another independent set of residuals and make its covariance matrix the same as $\sum_{\mathbf{r}}$.
3. Generate independent random variates **V**. The size of **V** is the same as the size of **X**.
4. Regress **V** on both **S** and **X** to generate another residuals set **b**, which is independent of both **S** and **X**.
5. Compute the covariance matrix of the second residuals set $\sum_{\mathbf{b}}$. Note that although the new set of residuals $\mathbf{b}$ is independent of $\mathbf{S}$, $\mathbf{X}$, and $\mathbf{r}$, the covariance matrix $\sum_{\mathbf{b}}$ is different than $\sum_{\mathbf{r}}$.
6. Compute a new residuals set $\mathbf{e}$ by scaling the (normalized) set of the independent residuals $\mathbf{b}$ to have the same covariance matrix as the covariance matrix $\sum_{\mathbf{r}}$ of original dataset:

$$\mathbf{e} = \left(\sum_{\mathbf{r}}\right)^{0.5} \left(\sum_{\mathbf{b}}\right)^{-0.5} \mathbf{b} \qquad (1)$$

7. Calculate the new perturbed attributes **Y**:

$$\mathbf{Y} = \mathbf{u} + \mathbf{e} \qquad (2)$$

Muralidhar and Sarathy [10] also proved that EGADP is optimal in terms of data utility and data security when all relationships in original *centralized* datasets are linear as in the case of multivariate normally-distributed datasets. In this study, we want to assess empirically the *scalability* of EGADP to the case of *distributed* horizontally-divided datasets in terms of optimality of data utility and data security.

3 Data Utility and Security Measures

Since the focus of this study is *linear* relationships, maintaining and reproducing aggregate measures such as mean, correlation matrix, and covariance matrix in masked datasets are adequate data utility measures [2, 8-10]. This is based on the *sufficient statistics theory* [1, 4, 7]. Therefore, if the mean vector, correlation matrix and covariance matrix of masked datasets are *similar* to the ones of original datasets, masked datasets can be utilized in ways similar to the uses of original datasets. For EGADP to be scalable to the case of distributed datasets in terms of its *data utility optimality*, these measures of masked datasets should *exactly* (not just *similarly*) match the measures of original datasets.

For data security, there are two conditions that should be satisfied [9]. Masked attributes $\mathbf{Y}$ should be independent of confidential attributes $\mathbf{X}$ given non-confidential attributes $\mathbf{S}$. In addition, $\mathbf{S}$ should be always a better predictor for $\mathbf{X}$ than $\mathbf{Y}$. Both mean that a snooper will always use $\mathbf{S}$ to obtain more accurate prediction results. However, if (s)he tries to combine $\mathbf{Y}$ with $\mathbf{S}$ to improve the prediction of $\mathbf{X}$, (s)he gains nothing (avoid partial and inferential disclosure). When all relationships are *linear* as in our motivation example, these conditions can be translated in terms of canonical correlation *CC* in the following equality and inequality [12]:

$$CC\left(\mathbf{X}|\mathbf{S}\right) = CC\left(\mathbf{X}|\mathbf{S},\mathbf{Y}\right) \qquad (3)$$

$$CC\left(\mathbf{X}|\mathbf{S}\right) \geq CC\left(\mathbf{X}|\mathbf{Y}\right) \qquad (4)$$

4 Masking Horizontally-Distributed Datasets Using EGADP

The procedure of masking horizontally-distributed datasets using EGADP is very simple and direct. Each party separately masks and perturbs his own original (sub-)dataset using EGADP as in the case of centralized datasets. Once this is done, each party can freely share their masked datasets. For this step, there are two possible and indifferent scenarios. In the first scenario, all masked datasets are sent to one specific (agreed on) party. The responsibility of this party is to combine all masked (sub-)datasets in one integrated masked dataset, and then forward it to all involved parties. In the second scenario, each party sends his masked dataset to all other parties. Then, each party builds his own integrated dataset. Now, each party can use the compiled dataset to build any model or run any statistical analysis s(he) wants without any restriction. Nevertheless, if the goal of sharing data is to build a specific model, all parties can cooperate in building that model using the integrated masked dataset.

Table 6: Masked Dataset Combining the Three Banks' Masked Datasets

	Masked Dataset*						
Customer	Non-Confidential Attributes (S)			Masked Attributes (Y)			Bank
No	S_1	S_2	S_3	Y_1	Y_2	Y_3	No
1	1	102.810	38.468	83.826	14.510	29.923	2
2	1	59.328	36.755	55.474	19.666	49.066	1
3	0	101.840	54.821	92.993	22.447	51.948	3
4	1	99.733	51.979	83.535	16.338	43.184	1
5	1	70.282	53.541	68.409	14.888	49.841	2
:	:	:	:	:	:	:	:
49,996	0	109.810	53.500	104.720	15.645	53.556	1
49,997	0	134.430	61.147	111.200	27.047	62.565	2
49,998	1	101.970	57.772	91.667	12.950	25.935	3
49,999	1	93.589	47.022	73.672	13.373	53.501	2
50,000	1	99.582	53.304	76.545	12.997	54.152	1

* The unit of S_2, S_3, Y_1, Y_2, Y_3 is \$000

5 Results and Discussion

In our motivation example, Bank 1, Bank 2 and Bank 3 mask independently their own confidential attributes **X** (X_1, X_2, X_3) using EGADP. This means that Bank 1, Bank 2 and Bank 3 perturb separately 18,000, 20,000 and 12,000 records, respectively. Then they freely share the masked datasets (S_1, S_2, S_3, Y_1, Y_2, Y_3) instead of their original confidential datasets (S_1, S_2, S_3, X_1, X_2, X_3). The three different datasets are combined in one dataset at one site and shared among all three banks. Table 6 lists few examples from this combined masked dataset (compare this table with Table 1). Each party, now, can build any model or run any statistical analysis from this integrated masked 50,000-records dataset.

When we compare the statistical measures of each party's dataset before masking (original dataset) with the corresponding statistical measures after masking, we find them identical (compare Table 7, Table 8, and Table 9 with Table 3, Table 4, and Table 5, respectively). This indicates that all banks could successfully apply EGADP on their own (centralized) datasets. Accordingly, all banks achieve (locally) optimal data utility and all original linear relationships were successfully reproduced in their masked datasets.

When we compare the statistical measures of the combined 50,000-records masked dataset with the statistical measures of the original combined dataset, we also find them identical (please refer to Table 10 and compare it with Table 2). This shows that EGADP is *scalable* in terms of the optimality of data utility in the case of horizontally-*distributed* datasets. This could be contributed to the additive nature of the statistical measures that EGADP tries to maintain. The three banks can now obtain the same results from the masked integrated dataset as with the original integrated dataset when they run different multivariate analyses or build models such as multiple linear regression models [10].

Let assume that Bank 1 wants to build the following multiple regression model from the original integrated 50,000-records dataset (i.e. **S**, **X**):

$$X_3 = \beta_0 + \beta_1 S_1 + \beta_2 S_2 + \beta_3 S_3 + \beta_4 X_1 + \beta_5 X_2 + \varepsilon$$

However, Bank 1 has only access to the masked 50,000-records dataset (i.e. **S**, **Y**). Thus, the bank will try to build the following model instead:

$$Y_3 = \alpha_0 + \alpha_1 S_1 + \alpha_2 S_2 + \alpha_3 S_3 + \alpha_4 Y_1 + \alpha_5 Y_2 + \varepsilon$$

The bank hopes that the regression coefficient estimated from the masked dataset would be as close as possible to the ones estimated from the original dataset (i.e. $\alpha_i \approx \beta_i$, i = 0,..,5). The bank built the regression model and estimated the regression coefficients from the masked dataset. When we compare these estimated coefficients with the ones estimated from the original dataset, we find that they are identical: $\alpha_0 = \beta_0 = 27.2069$, $\alpha_1 = \beta_1 = 0.1143$, $\alpha_2 = \beta_2 = 0.0063$, $\alpha_3 = \beta_3 = -0.1038$, $\alpha_4 = \beta_4 = 0.0315$, and $\alpha_5 = \beta_5 = 1.2390$. This exceeds the expectations of Bank 1.

This example demonstrates the advantage of the suggested approach of using EGADP to mask horizontally-distributed datasets: getting exact models from the masked data as with the original data. Another advantage of this approach is that the three banks are not limited to a specific analysis or model (e.g. a specific pre-defined dependent variable) as in the case of other protection methods (cf. Karr et al. [5]). Each bank can build one or more models that are different from the models built by other banks.

These perfect data utility results and flexibility mean nothing if the optimality of data security of EGADP does not scale well from *centralized* datasets to *distributed* datasets. For the individual three banks' datasets, the two (canonical correlation CC) data security conditions discussed earlier are

satisfied as expected (please refer to the first three lines in Table 11).

However, the main concern is that when the banks combine the three masked distributed datasets, a snooper can learn about confidential attributes more than what was intended before data release. By checking the two security conditions for the banks' 50,000-records integrated dataset, the optimality of EGADP data security seems scaling well in the case of horizontally-distributed datasets and the two security conditions are satisfied (please refer to the last line in Table 11).

The above discussion of both data utility and data security of the suggested approach shows that EGADP can be effectively applied to mask horizontally-distributed datasets when all the relationships in original datasets are linear or the goal is only to reproduce linear relationships in masked datasets.

6. Conclusion and Further Research

The results of this study prove that EGADP can be used effectively for masking horizontally-distributed datasets while enabling building accurate data mining models and obtaining accurate statistical analysis results. We are now working on developing the theoretical proofs for the results presented in this study and elaborating more on the possible uses of EGADP for protecting horizontally- and vertically- distributed datasets.

Table 7: Data Utility Measures (Statistical Measures) for Bank 1's Masked Dataset

Summary Statistics		**Non-Confidential Attributes S**			**Confidential Attributes Y**		
		Correlation					
Mean	**Attribute**	S_1	S_2	S_3	Y_1	Y_2	Y_3
0.50	S_1	1					
100.17	S_2	-0.00650	1				
50.07	S_3	-0.00511	0.700	1			
80.18	Y_1	-0.01097	0.798	0.748	1		
20.03	Y_2	0.00094	0.507	0.402	0.252	1	
50.08	Y_3	0.00292	0.304	0.204	0.151	0.604	1
Standard Deviation		**Covariance**					
	Attribute	S_1	S_2	S_3	Y_1	Y_2	Y_3
0.50	S_1	0.250					
19.94	S_2	-0.065	397.720				
10.02	S_3	-0.026	139.780	100.320			
19.91	Y_1	-0.109	317.040	149.270	396.570		
5.02	Y_2	0.002	50.778	20.194	25.195	25.201	
10.01	Y_3	0.015	60.599	20.419	30.087	30.362	100.120

Dataset Size: 18,000 records

Table 8: Data Utility Measures (Statistical Measures) for Bank 2's Masked Dataset

Summary Statistics		Non-Confidential Attributes S			Confidential Attributes Y		
		Correlation					
Mean	Attribute	S_1	S_2	S_3	Y_1	Y_2	Y_3
0.50	S_1	1					
99.90	S_2	0.00405	1				
49.99	S_3	0.00100	0.699	1			
80.01	Y_1	0.00212	0.802	0.752	1		
19.93	Y_2	0.00143	0.494	0.396	0.246	1	
49.88	Y_3	0.01529	0.295	0.190	0.143	0.598	1
Standard Deviation	Attribute	Covariance					
		S_1	S_2	S_3	Y_1	Y_2	Y_3
0.50	S_1	0.250					
20.11	S_2	0.041	404.380				
10.03	S_3	0.005	141.030	100.660			
20.16	Y_1	0.021	325.120	152.030	406.430		
4.99	Y_2	0.004	49.572	19.828	24.784	24.949	
9.99	Y_3	0.076	59.208	19.029	28.834	29.839	99.879

Dataset Size: 20,000 records

Table 9: Data Utility Measures (Statistical Measures) for Bank 3's Masked Dataset

Summary Statistics		Non-Confidential Attributes S			Confidential Attributes Y		
		Correlation					
Mean	Attribute	S_1	S_2	S_3	Y_1	Y_2	Y_3
0.50	S_1	1					
99.92	S_2	0.00216	1				
49.92	S_3	-0.00278	0.702	1			
79.71	Y_1	0.00075	0.799	0.750	1		
20.07	Y_2	0.00400	0.500	0.405	0.254	1	
50.08	Y_3	-0.00097	0.303	0.212	0.160	0.597	1
Standard Deviation	Attribute	Covariance					
		S_1	S_2	S_3	Y_1	Y_2	Y_3
0.50	S_1	0.250					
19.90	S_2	0.022	396.110				
9.92	S_3	-0.014	138.610	98.429			
19.86	Y_1	0.007	315.900	147.690	394.370		
4.98	Y_2	0.010	49.541	20.002	25.087	24.773	
10.00	Y_3	-0.005	60.396	20.990	31.823	29.707	100.000

Dataset Size: 12,000 records

Table 10: Data Utility Measures (Statistical Measures) for the Masked Three Banks' Datasets (Combined)

Summary Statistics		Non-Confidential Attributes S			Masked Attributes Y		
		Correlation					
Mean	Attribute	S_1	S_2	S_3	Y_1	Y_2	Y_3
0.50	S_1	1					
100.00	S_2	-0.00023	1				
50.00	S_3	-0.00214	0.700	1			
80.00	Y_1	-0.00295	0.800	0.750	1		
20.00	Y_2	0.00184	0.500	0.400	0.250	1	
50.00	Y_3	0.00689	0.300	0.200	0.150	0.600	1
Standard Deviation	Attribute	Covariance					
		S_1	S_2	S_3	Y_1	Y_2	Y_3
0.50	S_1	0.250					
20.00	S_2	-0.002	400.000				
10.00	S_3	-0.011	140.000	100.000			
20.00	Y_1	-0.029	320.000	150.000	400.000		
5.00	Y_2	0.005	50.000	20.000	25.000	25.000	
10.00	Y_3	0.034	60.000	20.000	30.000	30.000	100.000

Dataset Size: 50,000 records

Table 11: Data Security Measures Using Canonical Correlation

Dataset	Dataset Size	Condition 1			Condition 2		
		CC(X|S)	=	CC(X|S,Y)	CC(X|S)	≥	CC(X|Y)
Bank 1 Dataset	18,000	0.8940	=	0.8940	0.8940	≥	0.7992
Bank 2 Dataset	20,000	0.8949	=	0.8949	0.8949	≥	0.8008
Bank 3 Dataset	12,000	0.8927	=	0.8927	0.8927	≥	0.7969
Integrated Dataset	50,000	0.8940	=	0.8940	0.8940	≥	0.7993

7. References

[1] Anderson, T. W., *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Hoboken, N.J.: Wiley-Interscience, 2003.

[2] Burridge, J., "Information Preserving Statistical Obfuscation," *Statistics and Computing* (13:4), October 2003, pp. 321-327.

[3] Clifton, C., "Security and Privacy," in *The Handbook of Data Mining*, N. Ye (ed.), Mahwah, N.J.: Lawrence Erlbaum Associates, Publishers, 2003, pp. 441-452.

[4] Johnson, R. A. and Wichern, D. W., *Applied Multivariate Statistical Analysis*, 4th ed. Upper Saddle River, N.J.: Prentice Hall, 1998.

[5] Karr, A. F., Lin, X., Sanil, A. P., and Reiter, J. P., "Secure Regression on Distributed Databases," Retrieved April 28, 2008, from://www.niss.org/technicalreports/tr141.pdf.

[6] Kotz, S., Johnson, N. L., Balakrishnan, N., and Johnson, N. L., *Continuous Multivariate Distributions*, 2nd ed. New York: Wiley, 2000.

[7] Lehmann, E. L. and Casella, G., *Theory of Point Estimation*, 2nd ed. New York: Springer, 1998.

[8] Muralidhar, K., Parsa, R., and Sarathy, R., "A General Additive Data Perturbation Method for Database Security," *Management Science* (45:10), October 1999, pp. 1399-1415.

[9] Muralidhar, K. and Sarathy, R., "A Theoretical Basis for Perturbation Methods," *Statistics and Computing* (13:4), October 2003, pp. 329-335.

[10] Muralidhar, K. and Sarathy, R., "An Enhanced Data Perturbation Approach for Small Data Sets," *Decision Sciences* (36:3), August 2005, pp. 513-529.

[11] Sanil, A. P., Karr, A. F., Lin, X., and Reiter, J. P., "Privacy Preserving Regression Modelling via Distributed Computation," in *2004 ACM SIGKDD International Conference on Knowledge Discovery and Data*, Seattle, WA, USA, 2004, pp. 677-682.

[12] Sarathy, R. and Muralidhar, K., "The Security of Confidential Numerical Data in Databases," *Information Systems Research* (13:4), December 2002, pp. 389-403.