An Attack-Resistant Hybrid Data-Privatization Method with Low Information Loss

Kalpana Singh, Lynn Batten

An Attack-resistant Hybrid Data-privatization Method with Low Information Loss

Kalpana Singh, Lynn Batten
School of Information Technology, Deakin University, Melbourne, Australia
{kalpana, lynn.batten}@deakin.edu.au

Abstract. We examine a recent proposal for data privatization by testing it against well-known attacks; we show that all of these attacks successfully retrieve a relatively large (and unacceptable) portion of the original data. We then indicate how the data privatization method examined can be modified to assist it to withstand these attacks and compare the performance of the two approaches. We also show that the new method has better privacy and lower information loss than the former method.

Keywords. data privatization; information loss; Chebyshev polynomial; Spectral Filtering; Bayes-Estimated Data Reconstruction; data mining.

1 Introduction and Background

1.1 Data Privatization

Privacy preservation is an important issue in many data mining applications dealing with sensitive data such as health-care records. Privacy preserving data mining (PPDM) has become an important enabling technology for integrating data and determining interesting patterns from private collections of databases, thus improving productivity and competitiveness for many businesses. PPDM requires data modification which limits information loss (thus increasing utility) as it is intended that a legitimate receiver of the modified data be able to recover the original data needed for a response. Perturbation techniques have to manage the intrinsic trade-off between preserving data privacy and information loss, as each affects the other. Several perturbation techniques [1]-[5] have been proposed for mining purposes, but in all these papers, privacy and utility are not satisfactorily balanced. In the research literature, there are two general approaches to privacy preserving data mining: the randomization approach [1] and the secure multi-party computation approach [6]. We focus only on the former because it can distort data more efficiently than the latter.

There are two major randomization methods: Random Perturbation [2] and Randomized Response [5]. The former is a technique which deals mostly with numerical data, perturbing attribute by attribute, and concentrating on a statistical analysis of the data; it is a well-studied sanitization method that simultaneously allows access to the data by publishing them and at the same time preserving the privacy of the data. Randomized Response perturbs multiple attributes rather than one at a time, and so we ignore this method.

In the literature, perturbation is of two main types: additive [1], [2] and multiplicative [3], where random data (noise) is respectively either added or multiplied with the original data. As shown by Kargupta et al. multiplicative noise techniques can pro-

vide a good level of privacy (http://www.csee.umbc.edu/~hillol/PUBS/kargupta_privacy03a.pdf), while additive noise perturbation techniques are more effective in reducing information loss [2].

There is a growing body of literature on additive perturbation techniques which work by adding random noise to the data in such a way that the individual data values are distorted while, at the same time, the underlying distribution properties are preserved thus helping to reduce information loss. Agrawal and Srikant [1] proposed a scheme for PPDM using random perturbation in which a random number is added to the value of each sensitive attribute. It has been shown [1] that this scheme suffers from information loss, but Agrawal and Aggarwal [7] developed a novel reconstruction algorithm which minimizes the information loss of the former scheme. Liu has proposed a multiplicative method [3] which improves on the level of privacy achieved in [1] and [2], but with reduced utility.

This leads us to the following research question: ***Given a data-privatization method which leaks data under certain attacks, can it be improved to withstand these attacks without affecting information loss?***

In this paper, we demonstrate that this is possible by taking a particular example of a data-privatization technique, showing that it leaks data under three attacks and then adapting it to withstand these attacks while retaining low information loss characteristics.

1.2 The Research Literature

The usefulness of additive noise perturbation techniques in preserving privacy was firstly questioned by Kargupta et al. [8] who showed that attackers can derive a good estimation of the original dataset values from the perturbed dataset using a *spectral filter* that exploits some theoretical properties of random matrices and, as a result, the data privacy can be seriously compromised. Huang et al. [9] further proposed two data reconstruction algorithms which are efficient when the added noise is independent of the original data; one is based on Principal Component Analysis [9], the other one chooses Maximum Likelihood Estimation [9] to estimate the data.

The purpose of the current paper is to test a specific additive perturbation method (described in Section 2) to see how well it withstands three classical additive data-reconstruction attacks. We choose: Spectral Filtering (SPF) [8], Bayes-Estimated Data Reconstruction (BE-DR) [9] and Multiple Miner attack with Fusion (MDMF) [10]. We use the SPF method because it has a good track record in reconstructing original data based on additive perturbation; it is based on eigenvalues of a covariance matrix and the theory of random matrices [8]. We choose BE-DR for its ease of calculation and also because of its similarity to the calculations of SPF. The MDMF method is a combination of multiple data mining [10] and fusion techniques [10]; we use WEKA software [11] for data mining techniques in this method.

The particular data-privatization method [12] we test is based on Chebyshev polynomials of the first kind [13] which are explained in detail in Section 2. This method was developed recently by a group of researchers [12] but needs testing to see if it withstands the classical attacks. For testing, we derive 4500 test sets and each is tested for reconstruction using the three attack methods just described. We show that in all cases, some of the original data can be recovered. In addition, for each attack, we are able to specify how to modify the data privatization algorithm to make the data

resistant to the attack with the result that we produce a revised data privatization method and show that it is resistant to these attacks.

In Section 2, we present the Chebyshev-based data-privatization method as described in [12]; Section 3 presents the attacks on this method, and results of these attacks. Section 4 describes our revised version of the Chebyshev-based data-privatization method and we produce the results of three attacks on it and compare the performance with the former method. Section 5 discusses our results.

2 The Chebyshev Polynomial Perturbation Method

An additive perturbation technique based on Chebyshev polynomials was presented in [12] and in this section we describe it in detail. The Chebyshev polynomials (http://mathworld.wolfram.com/ChebyshevPolynomialoftheFirstKind.html) used in the paper [12] are said to be Chebyshev *of the first kind* as opposed to Chebyshev polynomials *of the second kind.* In this paper we use the expression '*Chebyshev Polynomial' to refer exclusively to the Chebyshev Polynomials $T_n(t)$ of the first kind.*

The authors of [12] propose an additive perturbation algorithm based on Chebyshev polynomials as described below in sub-section 2.1.

2.1 Chebyshev Data Perturbation Algorithm (CDP)

1. Data: In the paper [12] numerical type data for computation, such as the age of a patient which is common in health data, are considered. While the original data are commonly in matrix form, we store them in vector form. Section 2.2 provides notations. The original data set is referred to as *O* and the added noise data as *N*. These have the same size.

2. Setting parameters: *m:* the (square integer) number of entries in the original vector (or matrix).

n: the degree of the Chebyshev polynomial of the first kind, $n \geq 2$.

l: a positive integer divisor of *m*, $l > 1$.

3. Data perturbation: The initial values of the above parameters *m*, *n* and *l*, are assumed to be fixed.

a) Preparation: Derive the n^{th} degree Chebyshev polynomial T_n.

b) Division process: Divide the original data in vector form into intervals of length *l*; this results in $D = \frac{m}{l}$ intervals. Label the *D* intervals $t_1, t_2, \ldots t_D$. The first interval t_1 contains the original data $o_{11}, o_{21}, \ldots o_{l1}$; the second contains the next *l* elements $o_{l+1\,1}, o_{l+2\,1}, \ldots o_{2l\,1}$ and so on.

c) CDP data processing: In this step we generate noise to add to the original data. If an element o_{i1}, $1 \leq i \leq m$, of the original data set is in interval j, $1 \leq j \leq D$, then we add it to the corresponding element of the noise matrix defined as

$$pt(t_i)_1 = T_n\left(-1 + \frac{1}{n} + \frac{2\left(1-\frac{1}{n}\right)j}{l+1}\right) \text{ obtaining } \tilde{p}_{i1} = o_{i1} + pt(t_i)_1.$$

And so, the general equation describing this method is, $\tilde{P} = O + N$. (1)

4. Restoration: When $\tilde{p}_{i1}$ and $pt(t_i)_1$ are known, the data can easily be restored by using: $o_{i1} = \tilde{p}_{i1} - pt(t_i)_1$.

Our aim is, knowing only $\tilde{P}$ and not O, to reconstruct the original dataset values from $\tilde{P}$ thus breaking the privatization method of the paper [12]. For the reconstruction we use the three attacks mentioned in Part B of Section 1.

2.2 Summary of Symbols and Values Used

This sub-section provides the notation for sub-section 2.1 and subsequent parts of the paper.

Table 1. Summary of symbols and definitions

Symbols	Definitions
O	original dataset matrix
o_{ij} , $1 \le i,j \le \sqrt{m}$ and o_{i1}, $1 \le i \le m$	elements of the original dataset with matrix and vector indexation respectively
$\tilde{P}$	perturbed dataset matrix $\tilde{P} = O + N$
$\widetilde{p_{ij}}$, $1 \le i,j \le \sqrt{m}$ and $\widetilde{p_{i1}}$, $1 \le i \le m$	elements of the perturbed dataset with matrix and vector indexation respectively
N	noise dataset matrix
$pt(t_i)$j , $1 \le i,j \le \sqrt{m}$ and $pt(t_i)_1$, $1 \le i \le m$	elements of the noise dataset with matrix and vector indexation respectively
$\hat{O}$	estimated dataset matrix
$\widehat{o_{ij}}$, $1 \le i,j \le \sqrt{m}$ and $\widehat{o_{i1}}$, $1 \le i \le m$	elements of the estimated dataset with matrix and vector indexation respectively
m_i	i^{th} row of the matrix M
M_{cov}	covariance matrix of the matrix M
I	identity matrix
σ and σ^2	standard deviation and variance of the noise matrix elements
μ_M	mean vector of the matrix M
$[\lambda_{min}, \lambda_{max}]$	bounds for the eigenvalues of a matrix

3 Description and Results of the Reconstruction Attacks

In this section we explore the SPF, BE-DR and MDMF reconstruction methods and examine how well they estimate the original data. We use the same assumptions on data mentioned in [8] (SPF), [9] (BE-DR) and [10] (MDMF) as appropriate, and use the notation of Table 1. We generate 4500 matrices using the algorithm described in Section 2; 1500 of these were of size 400, 1500 of size 1600 and 1500 of size 6400. Because our data is stored in vector form, we are at liberty to decide on the matrix size for the reconstruction, and since we work heavily with eigenvalues, which are easy to produce from square matrices, we assume that O is square.

We use MATLAB R2009b [14] and WEKA 3.7.7 [11] for our experiment analysis. Using the Chebyshev Polynomial method described in Section 2 and the test matrices described above. (Note that because MATLAB [14] was used to generate and attack the vectors, we were restricted to the largest vector size it can handle, which is 6400 elements.) We implemented all data reconstruction methods on the 4500 datasets mentioned above. In each case, we obtained an estimation of the original data set which was then compared with the original data set in terms of the success measures described in the papers [8], [9] and [10]. In each case, although we tested (the same) 4500 matrices in each attack, we present the details of only one of our matrices – one which has 6400 entries.

3.1 Spectral Filtering Reconstruction Method

Test Example

i We calculate the eigenvalues of the covariance matrix of the perturbed matrix $\tilde{P}$ of the fixed example matrix *O*. Then we calculate $\lambda_{min} = 0$ and $\lambda_{max} = 0.4556$ from which we obtain those noisy eigenvalues $\widetilde{\lambda_i}$ which satisfy the inequality $\lambda_{min} \leq \widetilde{\lambda_i} \leq \lambda_{max}$., that is, which are in the range (0, 0.4556). The remaining eigenvalues should be those of *O*. Now from the SPF algorithm [8] we can calculate the eigenvalues of the covariance matrix of the estimated matrix. In order to obtain the eigenvalues of $\hat{O}$, we need only consider those above or equal to λ_{max} . To measure the success of the attack, we calculate the closeness of eigenvalues of $\hat{O}$ and eigenvalues of *O*. We obtain $|\lambda_{63} - \widehat{\lambda_{63}}|$=0, $|\lambda_{64} - \widehat{\lambda_{64}}|$=0,.....,$|\lambda_{80} - \widehat{\lambda_{80}}|$=0; all these differences are (very close to) zero, so an attacker can easily reconstruct the original data by using the SPF method.

ii To check the condition when values of reconstruction error increases as SNR [2] decreases, we give here only the average of reconstruction error values - 0.4384 while the value of SNR is 0.2408. So, in this case also we achieve a successful attack.

iii Lower and upper bound analysis, From [2], we calculate the lower bound to be 35.0716, and the upper bound to be 56.1346. Using [2], $\|\hat{O} - O\|_F$ = 37.4507. Since $37.4507 \geq 35.0716$ and $37.4507 \leq 56.1346$, both lower and upper bound conditions are satisfied.

iv We get RMSE [2] = 0.4384. The fact that $0.4384 < 1$ means our estimated dataset is not erroneous; so the attacker has breached the privacy.

In summary, the SPF method breached the privacy of the data-privatization method. In fact, SPF reconstruction is known to work well against additive techniques [8].

3.2 Bayes-Estimated Data Reconstruction Method

Test Example

i We calculate RMSE = 0.5187 <1; so the attacker has breached the privacy ([2]).

ii We calculate the reconstruction error corresponding to every element of the data set in vector form: $|o_{11} - \widehat{o_{11}}|$ = 0.015161, $|o_{21} - \widehat{o_{21}}|$ = 0.014837, $|o_{31} - \widehat{o_{31}}|$ = 0.014654 and so on until $|o_{6400\ 1} - \widehat{o_{6400\ 1}}|$ = 0.000035037. From the calculation we find that the attack has been successful ([9]).

3.3 Multiple Data Mining and Fusion Reconstruction Method

We use seven data mining algorithms $Ms, s = 1, \ldots, 7$. The following 7 miners, representing four different categories, were selected from the WEKA software package, version 3.7.7 [11], and used in the attack: *Function-based*: Simple linear regression (M1), *Meta*: CVParameterSelection (M2), Stacking (M3), Vote (M4); *Rule-based*: ZeroR (M5); *Tree-based*: DecisionStump (M6) and REPTree (M7).

Test Example

i We calculate the RMSE value to be 0.4033<1, and so privacy has been breached ([2]).

ii We evaluate the success of attack by calculating $d(\widehat{o_{wj}}, o_{wj}) < d(\widetilde{p_{wj}}, o_{wj})$ for all $1 \leq w, j \leq \sqrt{m}$ ([10]) and find that 5337 elements out of 6400 elements of the matrices

have satisfied the inequalities. In this case, the attacker has obtained 83.39% of the original data and failed to recover 16.61 % of it.

iii The added noise is $\sum_{o_{wj}} |\widetilde{p_{wj}} - o_{wj}|$ =116.1852; the remaining noise is $\sum_{o_{wj}} |\widehat{o_{wj}} - o_{wj}|$ = 31.5011. So, we get the ratio $\frac{remaining\ \ noise}{added\ \ noise}$ = 0.2711<1. Because this is close to zero, the attack has been successful ([10]).

4 Presentation of Proposed Method and Results of Attacks

4.1 Proposed Method

In this section, we propose a new hybrid data perturbation method with better utility preservation and privacy preservation than that described in Section 2. It has been pointed out that attacks which work on an additive data perturbation method will also work on a multiplicative data perturbation method as the latter can be logarithmically transformed into an additive data perturbation method [15]; hence we avoid a straightforward additive or multiplicative method.

While the noise matrix (*N*) is again generated using a Chebyshev polynomial of the first kind, our proposed method includes an orthogonal, rotation transformation matrix (*R*) and a translation matrix (*T*). *R* is added because rotation transformations preserve the utility of the most critical information for many classification models [3]. *T* is added to increase resilience to attack [16]. *I* is the $\sqrt{m}x\sqrt{m}$ identity matrix.

Our proposed hybrid perturbation is defined as follows:

1. Data: We use the format of sub-section 2.1.

2. Setting parameters: The values *m, n* and *l* are as in sub-section 2.1.

3. Data perturbation: Data preparation and division are as in sub-section 2.1, part 3.

Data processing: In this step we generate the orthogonal matrix *R* to multiply by the original data and the translation matrix *T* as well as the noise to add to the original data. The choice of the type of *R* is based on observations in [3] stating "random orthogonal transformation seems to be a good way to protect data's privacy while preserving its utility." And, for a legitimate user, "it is possible to re-identify the original data through a proper rotation." Also, as part of future work, the authors suggest that "The random projection-based technique may be even more powerful when used with some other geometric transformation techniques like … translation, and rotation." [16; p. 105] In addition, the author of [3] demonstrates that a rotation matrix preserves data privacy and quality. Indeed, our current hybrid proposal confirms this suggestion.

Producing R: We use MATLAB [14] to produce an orthogonal matrix [3] on input of the size m.

Producing T: Again, MATLAB produces a translation matrix on input of m.

Each element of the noise vector $pt(t_i)_1$ as generated in (c) of Part 3, sub-section 2.1, and

$$\tilde{P} = \frac{RO^2}{100} + O + T + N \quad (2)$$

where $\tilde{P}$, *O* and N are as in Table 1. We divide by 100 to bring the values of the perturbed matrix within the range of values of the original data.

We test this hybrid method with the three additive attacks used to test the earlier method described in Section 2. As in the earlier testing, for all tests we choose 4500 'original' matrices of three different sizes, but this time, derive the perturbed matrices from (2). The experiment followed the methods of Section 3 and the attacker fails

with respect to each of the reconstruction methods. Due to space limitations, we merely summarize the experimental results in Table 2 to three decimal places.

Table 2. Comparative analysis between additive method [12] and our proposed method

Reconstruction attack methods And information loss	Additive Method [12]						Our Proposed Method				
	Size of matrices →	400 size	1600 size	6400 size	Level of privacy (VoD)		400 size	1600 size	6400 size	Level of privacy (VoD)	
	Measurement factor ↓				E[D]	Var [D]				E[D]	Var [D]
SPF	SNR	0.213	0.230	0.241	0.203	0.151	34.247	35.65	36.59	48.44	16.235
	RMSE	0.314	0.386	0.438			42.364	46.434	47.758		
	PoS	80.28	78.63	76.86			0.115	0.091	0.042		
	PoF	19.72	21.37	23.14			99.89	99.91	99.96		
BE-DR	RMSE	0.387	0.469	0.519	0.496	0.023	38.93	42.03	43.95	44.65	16.12
	PoS	73.68	71.85	69.27			0.493	0.454	0.441		
	PoF	26.32	28.15	30.73			99.51	99.55	99.56		
MDMF	RMSE	0.369	0.389	0.403	0.505	0.061	40.01	44.49	45.78	46.35	16.66
	PoS	87.45	85.79	83.39			0.428	0.416	0.304		
	PoF	12.55	14.21	16.61			99.57	99.58	99.70		
	Relative Noise/ Added Noise	0.189	0.258	0.271			10.27	14.36	15.851		

5 Summary and Discussion

5.1 Level of Privacy and Information loss

First, we calculate success and failure of the attacker attack by attack, averaged over all the matrices of the same size and for each of the old additive method of Section 2 and the new proposal of Section 4. Table 2 indicates that the results do not depend on matrix size. Success (PoS) measures how much of O is reconstructed successfully by the attacker and is measured by $\frac{Number\ \ of\ elements\ \ (|\hat{O}-O|<|\tilde{P}-O|)}{number\ \ of\ elements\ \ of\ O}$ [17] as a percentage; the Failure (PoF) [17] is defined as 100 minus this value. The actual values are given in Table 2.

We also use a variance-of-difference (VoD) method [3] for measuring privacy in order to compare the two methods discussed in this paper. VoD measures privacy of matrices column-wise and we can calculate it with no knowledge of the original data. The perfect estimation will have zero mean and variance. We define VoD for all $1 \leq i \leq m$ as $D_i = \hat{o}'_{i1} - o_{i1}$, and calculate E[$D_i$] the mean of all D_i and var [D_i] the variance of all D_i. The average values are shown in Table 2. The fact that the mean and variance of the D_i are higher than 1 means that the attacker cannot estimate the original data; lower than 1 indicates a breach of privacy.

The level of *information loss* is measured by using a dissimilarity function [2] between the original dataset O and the perturbed dataset $\tilde{P}$. This dissimilarity function is

denoted by Diss (O, $\tilde{P}$) and lies in [0,1]. A value of Diss (O, $\tilde{P}$) near 0 denotes low information loss. We calculate average information loss for our proposed method as 0.004310 which is very low. We also calculate the dissimilarity function for the additive method [12] obtaining *Diss (O, $\tilde{P}$)* = 0.864534. So, average information loss for our proposed method is 0.4310% and for the additive method of [12] is 86.4534%. We conclude that our proposed method has achieved low information loss in comparison with the additive perturbation method of equation (1).

In summary, we have shown how a data perturbation method can be attacked by several reconstruction methods and then adjusted to withstand the attacks.

References

1. Agrawal, R., Srikant, R.: Privacy-Preserving Data Mining. In Proceedings of the ACM SIGMO Conference on Management of Data, Dallas, Texas, ACM Press. (2000) 439–450.
2. Datta, S.: On Random Additive Perturbation for Privacy Preserving Data Mining. Thesis report. (2004).
3. Liu, K.: Multiplicative Data Perturbation for Privacy Preserving Data Mining. Thesis report. (2007).
4. Singh, K., Zhong, J., Batten, L., Bertok, P.: An Efficient Solution for Privacy Preserving, Secure Remote Access to Sensitive Data. International conference of Advanced Computer Science and Information Technology. (2012) 173-191.
5. Du, W., Zhan, Z.: Using Randomized Response Techniques for Privacy-Preserving Data Mining. In Proceedings of The 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA. (2003) 505 – 510.
6. Du, W., Atallah, M.: Secure Multi-Party Computation Problems and Their Applications: A Review and Open Problems. In: New security paradigms workshop. (2001) 11–20.
7. Agrawal, D., Aggarwal, C.: On The Design and Quantification of Privacy Preserving Data Mining Algorithms. In Proccedings of the 20th ACM SIGACT-SIGMOD-SIGART Symposium on Principlesof Database Systems, California, USA. (2001) 247-255.
8. Kargupta, H., Datta, S., Wang, Q., Sivakumar, K.: On The Privacy Preserving Properties of Random Data Perturbation Techniques. In Proc. of the 3rd Int'l Conf. on Data Mining. (2003) 99–106.
9. Huang, Z., Du, W., Chen, B.: Deriving private information from randomized data. In Proceedings of the 2005 ACM SIGMOD Conference, Baltimroe, MD. (June 2005) 37–48.
10. Sramka, M., Safavi-Naini, R., Denzinger, J.: An Attack on The Privacy of Sanitized Data That Fuses the Outputs of Multiple Data Miners. In PADM. (2009) 130–137.
11. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H.: The WEKA Data Mining Software: An Update. SIGKDD Explorations, Volume 11, Issue 1. (2009).
12. Singh, K.: Recovering Private Data: A Comparision of Three Methods. Poster in proceedings of ATIS 2012. (2012) 24-25.
13. Mason, J.C., Handscomb, D.C.: Chebyshev Polynomials. Prentice Hall, SIAM, 2002.
14. Franco-Pereira, A.: An introductory course in MATLAB: MATLAB for beginners. Universidad Carlos III de Madrid. (2010). http://webs.uvigo.es/alba.franco/eng/Tutorial_completo.pdf.
15. Pandya, B., Singh, U., Bunkar, K., Dixit, K.: An Overview of Traditional Multiplicative Data Perturbation. International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 3. (2012) 424-429.
16. Liu, K., Kargupta, K., Ryan, J.: Random Projection-Based Multiplicative Data Perturbation for Privacy Preserving Distributed Data Mining. IEEE Transaction on Knowledge and Data Engineering, VOL. 18, NO. 1. (2006) 92-106.
17. McMullen, C.: Probability Theory. Harvard University. (2011) available at http://www.math.harvard.edu/~ctm/papers/home/text/class/harvard/154/course/course.pdf.