Volume 2, Issue 3, March 2012 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

An Overview of Traditional Multiplicative Data Perturbation

Bhupendra Pandya, Umesh Kumar Singh, Kamal Bunkar, Keerti Dixit
1.Institute of Computer Science, Vikram University
2.Institute of Computer Science, Vikram University
3.Institute of Computer Science, Vikram University
4.Institute of Computer Science, Vikram University
bhupendra20pandya@yahoo.co.in

Abstract: A Statistical database (SDB) is a database system that allows its users to retrieve aggregate statistics (e.g., sample mean and variance) for a subset of the entities represented in the database and prevents the collection of information on specific individuals. In the statistics community, there has been extensive research on the problem of securing SDBs against disclosure of confidential information. This is generally referred to as statistical disclosure control. Statistical disclosure control approaches suggested in the literature are classified into four general groups: conceptual, query restriction, output perturbation and data perturbation [1]. Conceptual approach provides a framework for better understanding and investigating the security problem of statistical database at the conceptual data model level. It does not provide a specific implementation procedure. The Query Restriction approach offers protection by either restricting the size of query set or controlling the overlap among successive queries. The Output Perturbation approach perturbs the answer to user queries while leaving the data in the database unchanged. The Data Perturbation approach introduces noise into the database and transforms it into another version. This paper primarily focuses on the data perturbation approaches.

Keyword: Multiplicative Data Perturbation

Introduction: Adding random noise to the private database is one common data perturbation approach. In this case, a random noise term is generated from a prescribed distribution, and the perturbed value takes the form: $y_{ij} = x_{ij} + r_{ij}$, where x_{ij} is the i^{th} attribute of

the j^{th} private data record, and r_{ij} is the corresponding random noise. In the statistics community, this approach was primarily used to provide summary statistical information (e.g., sum, mean, variance, etc.) without disclosing individual's confidential data. In the privacy preserving data mining area, this approach was considered [2,3] in for building decision tree classifiers from private data. Recently, many researchers have pointed out that additive noise can be easily filtered out in many cases that may lead to compromising the privacy [4,5]. Given the large body of existing signal-processing literature on filtered random additive noise, the utility of random additive noise for privacy-preserving data mining is not quite clear.

The Possible drawback of additive noise makes one wonder about the possibility of using multiplicative noise (i.e., $y_{ij} = x_{ij} * r_{ij}$) for protecting the privacy of the data. Two basic forms of multiplicative noise have been well studied in the statistics community [6]. One multiplies each data element by a random number that has a truncated Gaussian distribution with mean one and small variance. The other takes a logarithmic transformation of the data first, adds multivariate Gaussian noise, then takes the exponential function exp (.) of the noise-added data. As noted in the former perturbation scheme was once used by the Energy Information Administration in the U.S. Department of Energy to mask the heating and cooling degree days, denoted by x_{ij}. A random noise r_{ij} is generated from a Gaussian distribution with mean 1 and variance 0.0225. The random noise is further truncated such that the resulting number r_{ij} satisfies $0.01 \leq |r_{ij}-1| \leq 0.6$. The perturbed data $x_{ij}r_{ij}$ were released.

This paper gives a brief review of these two perturbation schemes.

Perturbation Scheme I:

Perturbation Scheme: Let x_i be the i^{th} attribute of a private database. Let x_i be the private value for the i^{th} attribute of the j^{th} record in the database, i = 1, . . . , n, j=1,

. . . , m. Let r_{ij} denote the random noise corresponding to x_{ij}. The perturbed data y_{ij} is

$y_{ij} = x_{ij} * r_{ij}$,

where r_{ij} is independent and identically chosen from a Gaussian distribution with mean 1 (usually $\mu_i = 1$) and variance σ_i^2. In other words, all r_{ij}'s for a given I follow the same distribution. In practice, the probability density of noise r (ignoring the subscript) is usually doubly truncated as follows:

$$f(r) = \frac{\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2\sigma^2}(r-\mu)^2)}{\frac{1}{\sqrt{2\pi}\sigma}\int_A^B \exp(-\frac{1}{2\sigma^2}(r-\mu)^2)dr} \quad \text{for } A < r < B.$$
$$= \frac{\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2\sigma^2}(r-\mu)^2)}{\Phi(\frac{B-\mu}{\sigma}) - \Phi(\frac{A-\mu}{\sigma})},$$

where A and B are the lower and upper truncation bounds and Ø(A) stands for the cumulative probability up to A. The above equation can be further simplified as

$$KZ(\frac{r-\mu}{\sigma}),$$

where $K = \frac{1}{\Phi(\frac{B-\mu}{\sigma}) - \Phi(\frac{A-\mu}{\sigma})}$, and $Z(x) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}x^2)$.

Statistical Properties of the Perturbed Data:

It has been proved [15] that the mean and variance of the original data attributes can be estimated from the mean and variance of the perturbed data.

Mean of x_i:

$$E(x_i) = \frac{E(y_i)}{\mu_i + K[Z(\frac{A-\mu_i}{\sigma_i}) - Z(\frac{B-\mu_i}{\sigma_i})]}.$$

Because the data owner will release μ_i, σ_i, A and B, the data receiver can compute the expected value of x_i.

Variance of x_i:

$Var(x_i) = E(x_i^2) - (E(x_i))^2$,

Where $E(x_i)$ can be easily calculated from the above equation and $(E(x_i))^2$
Can be computed from the follow equation:

$$\begin{aligned} Var[y_i] &= E(x_i^2)E(r_i^2) - (E(x_i)E(r_i))^2 \\ &= E(x_i^2)\{\sigma_i^2 + \mu_i^2 + \sigma_i^2 K[\frac{A-\mu_i}{\sigma_i}Z(\frac{A-\mu_i}{\sigma_i}) - \frac{B-\mu_i}{\sigma_i}Z(\frac{B-\mu_i}{\sigma_i})] \\ &\quad +2\sigma_i\mu_i K[Z(\frac{A-\mu_i}{\sigma_i}) - Z(\frac{B-\mu_i}{\sigma_i})]\} \\ &\quad -(E(x_i))^2\{\mu_i^2 + \sigma_i^2 K^2[Z(\frac{A-\mu_i}{\sigma_i}) - Z(\frac{B-\mu_i}{\sigma_i})]^2 \\ &\quad +2\sigma_i\mu_i K[Z(\frac{A-\mu_i}{\sigma_i}) - Z(\frac{B-\mu_i}{\sigma_i})]\}. \end{aligned}$$

Although the original attribute's mean and variance can be estimated from the perturbed data, the inner product and Euclidean distance among the data records are not necessarily preserved after perturbation. The following theorems depict this situation.

Theorem 5.1.3 Let $y_{ij} = x_{ij}\ r_{ij}$, where each r_{ij} is independent and identically chosen from a Guassian distribution with mean 1 and variance σ^2. Then

$$E(\sum_{i=1}^{n} y_{ij}y_{ik} - \sum_{i=1}^{n} x_{ij}x_{ik}) = 0;$$

$$Var(\sum_{i=1}^{n} y_{ij}y_{ik} - \sum_{i=1}^{n} x_{ij}x_{ik}) = \sigma^2 \sum_{i=1}^{n} x^2_{ij}x^2_{ik}$$

Proof:

$$E(\sum_{i=1}^{n} y_{ij}y_{ik} - \sum_{i=1}^{n} x_{ij}x_{ik}) = E(\sum_{i=1}^{n} x_{ij}r_{ij}x_{ik}r_{ik}) - \sum_{i=1}^{n} x_{ij}x_{ik}$$

$$= \sum_{i=1}^{n} E(x_{ij}r_{ij}x_{ik}r_{ik}) - \sum_{i=1}^{n} x_{ij}x_{ik}$$

$$= \sum_{i=1}^{n} x_{ij}E(r_{ij})x_{ik}E(r_{ik}) - \sum_{i=1}^{n} x_{ij}x_{ik}$$

$$= 0.$$

n n n

$$\mathrm{Var}\Big(\sum_{i=1} y_{ij}y_{ik} - \sum_{i=1} x_{ij}x_{ik}\Big) = \mathrm{Var}\Big(\sum_{i=1} x_{ij}r_{ij}x_{ik}r_{ik}\Big)$$

$$= \sum_{i=1}^{n} \mathrm{Var}(x_{ij}r_{ij}x_{ik}r_{ik}) + 2\sum_{p=1}^{n-1}\sum_{q=p+1}^{n} \mathrm{Cov}\,(x_{pj}r_{pj}x_{pk}r_{pk},\, x_{qj}r_{qj}x_{qk}r_{qk})$$

$$= \sum_{i=1}^{n} \mathrm{Var}(x_{ij}r_{ij}x_{ik}r_{ik})$$

$$= \sum_{i=1}^{n} \{E(x^2_{ij}r^2_{ij}x^2_{ik}r^2_{ik}) - (E(x_{ij}r_{ij}x_{ik}r_{ik}))^2\}$$

$$= \sum_{i=1}^{n} \{(1+\sigma^2)\, x^2_{ij}x^2_{ik} - x^2_{ij}x^2_{ik}\}$$

$$= \sigma^2 \sum_{i=1}^{n} x^2_{ij}x^2_{ik}.$$

The above theorem shows that although the inner product is preserved on expectation, the variance of the error could be very large.

Theorem 5.1.4 Let $y_{ij} = x_{ij}\, r_{ij}$, where each r_{ij} is independent and identically chosen from a Gaussian distribution with mean 1 and variance σ^2. Then

$$E\Big(\sum_{i=1}^{n}(y_{ij} - y_{ik})^2 - \sum_{i=1}^{n}(x_{ij} - x_{ik})^2\Big) = \sum_{i=1}^{n} \sigma^2\,(x^2_{ij} + x^2_{ik})$$

Proof: Let LHS denotes the left hand side of the above equation. We have

$$\mathrm{LHS} = E\Big(\sum_{i=1}^{n}(x_{ij}r_{ij} - x_{ik}r_{ik})^2\Big) - \sum_{i=1}^{n}(x_{ij} - x_{ik})^2$$

$$= E\Big(\sum_{i=1}^{n}(x^2_{ij}r^2_{ij} + x^2_{ik}r^2_{ik} - 2\, x_{ij}r_{ij}x_{ik}r_{ik})\Big) - \sum_{i=1}^{n}(x_{ij} - x_{ik})^2$$

$$= \sum_{i=1}^{n}((1+\sigma^2)x^2_{ij} + (1+\sigma^2)\, x^2_{ik} - 2\, x_{ij}x_{ik})) - \sum_{i=1}^{n}(x_{ij} - x_{ik})^2$$

$$= \sum_{i=1}^{n}((x_{ij} - x_{ik})^2 + \sigma^2\,(x^2_{ij}+x^2_{ik})) - \sum_{i=1}^{n}(x_{ij} - x_{ik})^2$$

$$= \sum_{i=1}^{n}(x_{ij} - x_{ik})^2 + \sum_{i=1}^{n}\sigma^2\,(x^2_{ij} + x^2_{ik}) - \sum_{i=1}^{n}(x_{ij} - x_{ik})^2$$

$$= \sum_{i=1}^{n} \sigma^2\,(x^2_{ij} + x^2_{ik}).$$

The above theorem shows that the Euclidean distance is not preserved after perturbation.

Perturbation Scheme II:

Perturbation Scheme:

Let x_{ij} be the value for the i-th attribute of the j-th record in the database as before i=1... n,j=1…m. Let

We generate the random noise following the multivariate Gaussian Distribution N (0, c $\sum$ u), where 0 < c < 1 and $\sum$ u is the covariance matrix of variables u_1, u_2… u_n. We denote the noise as e_{ij}. Let

$$z_{ij} = u_{ij} + e_{ij},$$
$$y_{ij} = \exp(z_{ij})$$
$$= \exp(\mathrm{In}\, x_{ij} + e_{ij})$$
$$= x_{ij}\exp(e_{ij})$$
$$= x_{ij}\, h_{ij}.$$

This perturbed data y_{ij} is released then. Note scheme assumes that all x_{ij} are positive.

Statistical Properties of the Perturbed Data:

It has been proved [15] that the mean, variance and covariance of the original data attributes can be estimated from the perturbed data.

Mean of x_i: Let $\sigma_i^2 = c\,\mathrm{Var}(\mathrm{In}\, x_i)$. We have

$$E(x_i) = E(y_i) / \exp(\sigma_i^2/2)$$

Variance of x_i:

$Var(x_i) = E(x_i^2) - (E(x_i))^2$

$=(Var(u_i)/\exp(2\sigma_i^2)) - (E(x_i)^2/\exp(\sigma_i^2)) - (E(x_i))^2$

Covariance of x_i and x_j:

$$Cov(x_i, x_j) = \left\{ \frac{\sum_{k=1}^{m} y_{ik} y_{jk}}{\exp[(\sigma_i^2 + 2\rho\sigma_i\sigma_j + \sigma_j^2)/2]} - \frac{m \frac{\sum_{k=1}^{m} y_{ik}}{m} \frac{\sum_{k=1}^{m} y_{jk}}{m}}{\exp[\sigma_i^2 + \sigma_j^2]} \right\} / (m-1),$$

where ρ is the correlation coefficient of x_i and x_j, and it can be obtained from the perturbed data. Because the noise was generated to maintain the same correlation structure, the correlation between the perturbed data will be on average the same as that between the original data in log-scale.

Similar to perturbed scheme I, the inner product and Euclidean distance among the data records are not preserved after perturbation. The following theorem depicts this situation.

Privacy Issue:

On the surface, multiplicative perturbation seems to change the data more than additive perturbation. For example, perturbing a salary of $100,000 by adding $5000 (5% relative change) would be considered a compromise while at the same time perturbing a salary of $10,000 by adding $5000 (50% relative change) would preserve the privacy of the data. On the other hand, perturbing $100,000 and $10,000 by multiplying by 2 would be accepted because both have 100% relative change. However, by taking logarithms on the perturbed data, the multiplicative perturbation turns into an additive perturbation. Most specifically, for perturbation scheme I, the logarithmic transformation of y_{ij} gives us $\ln x_{ij} + \ln (r_{ij})$, where the noise term ln (rij) is chosen independent and identically from some distribution.

For perturbation scheme II, after logarithmic transformation, we have $\ln x_{ij} + e_{ij}$. The noise term is chosen from N (0, c $\sum_{\ln} X$), where $\sum_{\ln} X$ is the covariance of the original data in log scale. As noted in [5, 7], the privacy of the former "additive perturbation scheme" can be easily breached in many cases. The latter "additive perturbation scheme" generates random noise with similar covariance structure with the original data (in log scale), and therefore offers better privacy protection. This kind of perturbation has also been extensively investigated in the literature. In particular, the work in [7] shows that the accuracy of attacker's estimation of the original data gets worse as the similarity increases.

Before concluding this subsection, it should be noted that, traditionally, the privacy, denoted by ρ, provided by a perturbation techniques for continuous data is measured as the variance of difference between the original data and perturbed data [8], that is, Var(X-Y). Where X represent the original data attribute and Y the perturbed attribute. This measured can be made scale invariant with respect to the variance of X as ρ = Var(X-Y)/Var(X). This measure is suited to quantifying the privacy of the single attribute. In practice an attacker may also attempt to use a linear combination of the perturbed attributes to estimate confidential information of the linear combination the original attributes. Measuring the privacy offered for a linear combination is difficult because there are too many such combinations. A canonical correlation –based metric is used in [8], that can measure the maximum proportion of various that an attacker can explain for any linear combination of the original attributes, using a linear combination of the perturbed and non-confidential attributes. Let λ denotes the largest eigen value of the following matrix $c_{xx}^{-1} c_{xy} c_{xy}^{-1} c_{yx}$ Where c_{xx} denotes the covariance of X, c_{xy} the covariance of X and Y. The value of λ represents the maximum proportion of variability in any linear combination of X that can be explained by any linear combination of Y. The privacy is defined as ρ=1-λ. Thus for any linear combination of X at least 1- λ proportion of variability will remain unexplained.

These metrics do provide the data owner with meaningful information regarding the effectiveness of the perturbation method in some way. However they do not offer an insight on how the attackers could attack the perturbation if they had some prior knowledge about the data tried to address this issue by developing a Bayesian attacker model to assess the performance of the perturbation techniques on continuous micro-data. They specifically investigated the combination of both additive noise and multiplicative noise and allowed the attacker to use external data to enhance the chances of disclosing the identity of a target individual. Their simulation showed that the probability of the identity disclosure is a function of many key parameters like the variability amongst profiles in the original data, the amount of attacker's prior information, the amount of noise introduced in the data, etc.

Conclusion:

This paper briefly reviews two traditional multiplicative data perturbation techniques that have been well studied in the statistics community. These perturbations are primarily used to mask the private data while allowing summary statistics (e.g., sum, mean, variance and covariance) of the original data to be estimated.

In summary these multiplicative perturbations have the following advantages and disadvantages.

The multiplicative perturbation is relative, that is, large values in the original data are perturbed more than smaller value.

In practice, the first perturbation scheme is good if the data disseminator only wants to make minor changes to the original data; the second scheme assures higher security than the first one but maintains the data utility in the log-scale.

These perturbation schemes are equivalent to additive perturbation after the logarithmic transformation. Due to the large volume of research in deriving private information from the additive noise perturbed data, the security of these perturbation schemes is questionable.

The objective of these perturbation schemes is to mask the private data while allowing summary statistics to be estimated. However, problems in data mining are somewhat different. Data mining techniques, such as clustering, classification, prediction and association rule mining, are essentially relying on more sophisticated relationships among data records or data attributes, but not simple summary statistics. The traditional multiplicative perturbations distort each data element independently, therefore Euclidean distance and inner product among data records are usually not preserved, and the perturbed data can not be used for many data mining applications.

REFERENCES

[1] N. R. Adam and J. C. Worthmann, "Security-control methods for statistical databases: a comparative study," *ACM Computing Surveys (CSUR)*, vol. 21, 145-146 no. 4, pp. 515–556, 1989. [Online]. Available:http://portal.acm.org/citation.cfm?id=76895

[2] R. Agrawal and R. Srikant, "Privacy preserving data mining," in *Proceedings of the ACM SIGMOD Conference on Management of Data*, Dallas, TX, May 2000, pp. 439–450.

[3] D. Agrawal and C. C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms," in *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of Database Systems*, Santa Barbara, CA, 2001, pp. 247–255. [Online]. Available: http://portal.acm.org/citation.cfm?id=375602

[4] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, "On the privacy preserving properties of random data perturbation techniques," in *Proceedings of the IEEE International Conference on Data Mining*, Melbourne, FL, November 2003.

[5] S. Guo and X. Wu, "On the use of spectral filtering for privacy preserving data mining," in *Proceedings of the 21st ACM Symposium on Applied Computing*, Dijon, France, April 2006, pp. 622–626.

[6] J. J. Kim and W. E. Winkler, "Multiplicative noise for masking continuous data," Statistical Research Division, U.S. Bureau of the Census, Washington D.C., Tech. Rep. Statistics #2003-01, April 2003. 1963, ch. XII, pp. 213–217.

[7] Z. Huang, W. Du, and B. Chen, "Deriving private information from randomized data," in *Proceedings of the 2005 ACM SIGMOD Conference*, Baltimroe, MD, June 2005, pp. 37–48.

[8] K. Muralidhar, R. Parsa, and R. Sarathy, "A general additive data perturbation method for database security," *Management Science*, vol. 45, no. 10, pp. 1399–1415, 1999.

