Working Paper No. 27
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint ECE/Eurostat work session on statistical data confidentiality
(Luxembourg, 7-9 April 2003)

Topic (i): New theories and emerging methods

NEW DEVELOPMENTS IN PERTURBING NUMERICAL MICRODATA

Contributed Paper

Submitted by the University of Kentucky and Oklahoma State University, United States[1]

[1] Prepared by Krish Muralidhar (krishm@uky.edu) and Rathindra Sarathy (sarathy@okstate.edu).

Recent Research Results on the Conditional Distribution Approach for Data Perturbation

Krish Muralidhar
Gatton Research Professor
Gatton College of Business & Economics
University Of Kentucky
Lexington KY 40506-0034
krishm@uky.edu

Rathindra Sarathy
Associate Professor
Department of Management Science & Information Systems
Oklahoma State University
Stillwater OK 74078
sarathy@okstate.edu

Recent Research Results on the Conditional Distribution Approach for Data Perturbation

In this extended abstract, we provide a summary of our recent research on developing a theoretical basis for perturbation methods. We propose that, theoretically, generating perturbed values of the confidential variables from the conditional distribution of the confidential variables given the non-confidential variables, but independent of the original confidential variables. We show that if the perturbed values are generated from this approach, the resulting perturbed values have the same statistical characteristics as the original confidential variables, and maintain all relationships among the variables to be the same after perturbation as before perturbation. Furthermore, since given the non-confidential variables, the perturbed variables are independent of the original confidential variables, this method also provides intruders with no knowledge gain. For a complete description, please see Muralidhar and Sarathy (2003). In the following sections, we describe our efforts in developing techniques based on this theoretical approach for numerical, confidential variables.

Our initial effort focused on the desire to improve the performance of existing additive noise techniques for numerical, confidential variables. One of the key aspects of noise addition techniques was that all techniques assumed the basic noise addition model, of the form,

$$\mathbf{Y} = \mathbf{X} + \boldsymbol{\varepsilon}. \qquad (1)$$

Starting with the original proposal by Traub et al. (1984), researchers had provided several improvements to this procedure (Kim 1986, Sullivan 1989, Sullivan and Fuller 1989, Tendick 1991, Tendick and Matloff 1994). Most of these procedures use different statistical characteristics for $\boldsymbol{\varepsilon}$, in an attempt to satisfy disclosure risk and data utility requirements. The procedures suggested by Kim (1986) and Tendick and Matloff (1994) ensured that, when the underlying distribution was normal, the marginal distribution and covariance matrix of the perturbed variables were the same as that of the original, confidential variables. Other researchers (such as Fuller 1993, Tendick 1992) had also addressed the performance of this procedure in terms of disclosure risk. Yet there were considerable deficiencies in the noise-addition approach shown in equation (1).

General Additive Data Perturbation Method

At the time we started our research on this topic, one key component that was missing in approaches in prior studies was the treatment of the non-confidential variables. From equation (1) it is easy to see that the non-confidential variables play no role in this approach. Most prior studies assumed that all variables were confidential and hence had to be perturbed and that there were no non-confidential variables present. In our opinion, this was a rather restrictive assumption. By their nature, non-confidential variables may be available in their original form through other sources. The releasing agency, because of concerns about re-identification may choose to perturb non-confidential numerical variables. However, when these values are available from other sources in original form,

they can be used to compromise the values of confidential variables, resulting in *predictive disclosure* (as opposed to *identity disclosure*). Even if we assume that, because of re-identification risk, all *numerical* variables were to be perturbed, it leaves open the issue of *categorical* non-confidential variables. The noise addition approach is not suited for categorical variables and hence could not be used to perturb such variables. Thus, if there were non-confidential variables (either numerical or categorical) present, prior approaches essentially ignored the categorical variables in performing the perturbation.

Ignoring non-confidential variables during perturbation has obvious disadvantages. It is easy to show that, if ρ represents the correlation between a confidential and non-confidential variable and the if the variance of ε was d^2, after perturbation, the correlation between the perturbed and the non-confidential variables was $\rho/(1+d^2)^{0.5}$. In other words, regardless of the level of noise added, analysis of the released data for this relationship would provide different results compared to the original data. Thus, even if the perturbed confidential variables maintained the same covariance structure as the original confidential variables, their relationship with the non-confidential variables would be biased.

Our initial approach was to develop a technique that ensured that this does not occur. Our original study (Muralidhar et al. 1999) considers the entire set of variables (confidential (**X**), non-confidential (**S**), and perturbed (**Y**)) as having a joint (multivariate normal) distribution. Note that, prior to generating the perturbed variables, while we do not know the *individual values* of **Y**, we know the *desired joint distribution of* **X**, **S**, and **Y**. In order to maximize utility, it is desirable that the distribution of **Y** should be the same as that of **X**, and that the joint distribution of (**Y** and **S**) should be the same as that of (**X** and **S**). Further, in order to reduce disclosure risk, we specified that the covariance between the original and perturbed variables should be ($\lambda \times$ covariance of the original confidential variables), where λ represents the square of the first canonical correlation between **X** and **S**. This ensured that, for any linear combination of the confidential variables, the proportion of variability explained was no more than λ. We also showed that all previously proposed methods of noise addition were special cases of this approach. Hence, we called this approach the General Additive Data Perturbation (GADP) method. For complete details of GADP and illustrations, please refer to Muralidhar et al. (1999). The paper also shows that even when the underlying distribution of the variables is *not* multivariate normal, GADP is capable of maintaining the covariance matrix of **Y** to be the same as that of **X**, and the covariance between (**Y** and **S**) to be the same as that between (**X** and **S**). However, the marginal distribution of **Y** was considerably different from that of **X**. GADP is also very effective when the non-confidential variables are categorical.

Our research was further spurred by a desire to improve the disclosure risk characteristics of GADP. While GADP performed better than the other methods in terms of disclosure risk, it did not satisfy the strict requirements that are described in Dalenius (1977) and Duncan and Lambert (1986). Both these papers evaluate risk by comparing the information that is available prior to "data" release and post "data" release and contend that disclosure risk occurs when the intruder "gains" information from the released data.

The specific objective of perturbation methods is to provide access to microdata values of the confidential variables. Hence, these methods essentially dictated that there should be no information and knowledge gain (or correspondingly, a decrease in uncertainty) when microdata is released. One of the difficulties with these definitions is that it is extremely difficult to assess the extent of knowledge that an intruder possesses prior to microdata release. However, we can model the knowledge of the intruder based on the assumption that the intruder is intelligent and specifically intends to compromise the data, and that the intruder has verifiable, accurate, information of her/his own. Further, since agencies have to at least consider releasing aggregate data[1], it is reasonable to include this knowledge as part of the snooper's prior knowledge in compromising the data.

This model of the intruder can be considered as the "worst case" scenario, and has been used by several respected authors in the area (see for example, Fuller 1993, Fienberg 1997, Willenborg and de Waal 2001, Yancey et al. 2002). From an agency perspective, modeling the intruder in this manner has the significant advantage that if it can be assured that risk of disclosure is very low for such an intruder, it will be even lower for other (not so intelligent) intruders. Thus, we adopt the "worst case" scenario or the "determined intruder" in evaluating disclosure risk.

If the objective of perturbation is to defeat such an informed intruder, it is necessary that *providing access to microdata does not result in providing the intruder with additional information*. Further, since we assume that the intruder could have their own data and/or aggregate data has been provided, we can assume, from the perspective of value disclosure, that prior to microdata release the intruder has information on $R^2_{X|S}$. In order to provide no additional information when microdata is released, it is necessary that:

$$R^2_{X|S,Y} = R^2_{X|S}. \qquad (2)$$

In Muralidhar et al. (2001), we derived the conditions that are necessary in order to satisfy this (disclosure risk) requirement and simultaneously satisfy the previously defined data utility requirements (namely, distribution of **Y** is the same as that of **X**, and the joint distribution of (**Y** and **S**) is the same as that of (**X** and **S**)) for the multivariate normal distribution. The resulting expression was of the form:

$$\mathbf{Y} = \boldsymbol{\mu}_X + \boldsymbol{\Sigma}_{XS}\boldsymbol{\Sigma}_{SS}^{-1}(\mathbf{s}_i - \boldsymbol{\mu}_S) + \boldsymbol{\varepsilon}, \qquad (3)$$

where, $\boldsymbol{\mu}_X$ and $\boldsymbol{\mu}_S$ are the mean vectors of **X** and **S**, $\boldsymbol{\Sigma}_{XX}$, $\boldsymbol{\Sigma}_{SS}$, and $\boldsymbol{\Sigma}_{XS}$ represent the covariance of **X**, **S**, and between (**X** and **S**), respectively. Finally, $\boldsymbol{\varepsilon}$ has a multivariate normal distribution with mean vector **0** and covariance matrix, $\boldsymbol{\Sigma}_{\varepsilon\varepsilon} = \boldsymbol{\Sigma}_{XX} - \boldsymbol{\Sigma}_{XS}\boldsymbol{\Sigma}_{SS}^{-1}\boldsymbol{\Sigma}_{SX}$. From these specifications, it is easy to see that:

[1] If the release of data, even in aggregate form, leads to unacceptable disclosure risks, the agency may not even release the data, and therefore would suffer no liability from an intruder's knowledge gain. Hence, the concept of an intruder, in this case, is irrelevant.

$\mathbf{y_i} \sim f(\mathbf{X} \mid \mathbf{S} = \mathbf{s_i})$. (4)

In other words, in order to satisfy the dual requirements of disclosure risk and data utility in the multivariate normal case, it is necessary to first derive f(**X**|**S**) using the available data and then to generate the values of **Y** using this conditional distribution and the values of **S** and independent of **X**. Since given **S**, the values of **X** and **Y** are independent, equation (2) is satisfied. This implies that this approach does not provide additional information to an intruder. In addition, it can be easily shown that the distribution of **Y** is the same as that of **X** and that the joint distribution of (**Y** and **S**) is the same as that of (**X** and **S**). Note that, in the absence of **S**, this procedure also suggests that the perturbed values should be generated from a multivariate normal distribution that has the same characteristics as **X**, but independent of **X**. Thus, for the multivariate normal distribution, GADP provides an "optimal" solution to the perturbation problem by minimizing disclosure risk and maximizing data utility.

Note that, although we originally started out defining disclosure risk by using the R^2 measure, the perturbed values resulting from GADP prevent *not only value disclosure but also identity disclosure*. Theoretical proof for this statement can be found in Muralidhar and Sarathy (2003). However, this result can also be explained in simple terms as follows. The predictive distribution of an intruder is based on f(**X**|**S**). We have shown above and in Muralidhar and Sarathy (2003), that f(**X**|**S**,**Y**) = f(**X**|**S**). Consequently, regardless of the predictive measure, disclosure risk will be minimized.

An extension of the simple noise addition approach given in equation (1) is the more recent "model-based" approach. While Muralidhar et al. (1999, 2001) do not directly address the model-based approach, it is easy to show its relationship with GADP. In general model based approaches suggest that the perturbed values **Y** should be generated from a general model of the form:

$$\mathbf{Y} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \mathbf{S} + \boldsymbol{\beta}_2 \mathbf{X} + \boldsymbol{\varepsilon}. \quad (5)$$

When $\boldsymbol{\beta}_0, \boldsymbol{\beta}_1 = \mathbf{0}$, and $\boldsymbol{\beta}_2 = 1$, equation (5) reduces to the simple noise-addition approach in equation (1). We can also easily show that, in order to achieve the disclosure risk and data utility requirements, it is necessary that $\boldsymbol{\beta}_2 = \mathbf{0}$, $\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 = \boldsymbol{\mu}_X + \boldsymbol{\Sigma}_{XS}\boldsymbol{\Sigma}_{SS}^{-1}$, and

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0},\ \boldsymbol{\Sigma}_{XX} - \boldsymbol{\Sigma}_{XS}\boldsymbol{\Sigma}_{SS}^{-1}\boldsymbol{\Sigma}_{SX}) \quad (6)$$

These are the same specifications used in GADP. Thus, GADP represents the "optimal" model based approach, and the *only "model-based" approach* that will satisfy both the data utility and disclosure risk requirements.

It is important to note that we use this notion of prior information *only* to define the important concept of minimum disclosure risk. Some critics have argued that we should compare the disclosure risk resulting from our method to those of other techniques. In our opinion, this is relatively straightforward. For multivariate normal distributions, the

perturbed values are independent realizations from the true conditional distribution of **X**|**S**. Hence, in these cases, given **S**, **X** and **Y** are *statistically* independent. Statistical independence automatically implies that for any method of prediction, given **S**, the perturbed values **Y** provide no information regarding **X**. None of the other methods of additive perturbation or model-based approaches can guarantee this result. As an illustration, consider the simple case where there is a single confidential variable (X) and a single non-confidential (S) variable with a joint standard multivariate normal distribution with correlation ρ. Let the perturbed variable be Y. Let disclosure risk be measured by R^2, the proportion of variability in X that is explained by the S and Y. We can easily show that if additive noise is used, the resulting risk of disclosure is:

$$R^2_{X|\,S,Y} = \rho^2 + \frac{(1-\rho^2)^2}{(1-\rho^2)+d^2}.$$

where d^2 represents the variance of the noise term ε. When a model based approach of the form $Y = a + bX + cS + \varepsilon$, is used, the resulting risk of disclosure is:

$$R^2_{X|\,S,Y} = \rho^2 + \frac{b^2(1-\rho^2)^2}{b^2(1-\rho^2)+d^2}.$$

When GADP is used to perturb X, the resulting risk of disclosure is:

$$R^2_{X|\,S,Y} = \rho^2.$$

It is easy to verify that GADP provides the lowest risk of disclosure and is equal to an intruder's prior estimate of X using S alone. Our use of prior information possessed by an intruder essentially negates the need to measure disclosure risk in every case for every situation and every measure of disclosure.

Finally, even in cases where the joint distribution of the variables is not multivariate normal, GADP can be used effectively in certain situations. One such situation occurs when all the non-confidential variables are categorical (and hence can be represented by binary variables) and the impact of these variables is a simple mean shift in the confidential variables. In such situations, GADP provides the same level of data utility and disclosure risk as it does for variables with a multivariate normal distribution. Even in other cases, if the statistic of interest is the covariance matrix and if it can be assumed that all predictions will be based on the covariance matrix, GADP can be used effectively. However, in these cases, GADP would not provide the same level of data utility since the marginal distribution of the perturbed variables will be different from that of the original variables and any non-linear relationships will be distorted. Finally, in such situations, if a snooper employs non-linear models to estimate the confidential variables, it is possible that additional information is provided.

Copula-Based GADP Method

One of the key concerns of using GADP is the fact that when the individual variables are not normally distributed, the marginal distribution of the perturbed values is different from that of the original values. While the conditional distribution approach is theoretically sound, implementing the conditional distribution approach for non-normal multivariate distributions is significantly more complicated than it is for the multivariate normal distribution. Even characterizing the joint distribution when the individual variables can take on arbitrary marginal distributions is very difficult. To derive the conditional distribution in these cases may not even be mathematically feasible. This required us to consider alternatives that would provide us with the ability to approximate the joint distribution when the individual marginal distributions are not normal.

Copulas offer just such an approximation. Copulas are often used in the statistical literature as a method for joining distributions with arbitrary marginal distributions. A variety of copulas have been proposed in the literature. One such copula is the multivariate normal copula. We employed the multivariate normal copula to generate conditional realizations for non-normal datasets (referred to as C-GADP). For a complete description of this method, please refer to Sarathy et al. (2002). The copula method can be described as follows:

1. Identify the marginal distribution of attributes $X_1,\ldots,X_n$, $S_1, \ldots S_m$.
2. Compute pair-wise **rank** order correlation matrix (**R**) of the original database.
3. Compute product moment correlation matrix $\boldsymbol{\rho}$ using **R** using the transformation $\rho_{ij} = 2\times\text{Sin}(\pi \times r_{ij}/6)$ (see Kruskal 1958).
4. Compute the new variables $\mathbf{X}^*$ and $\mathbf{S}^*$.
5. Apply GADP to variables $\mathbf{X}^*$ and $\mathbf{S}^*$ to generate $\mathbf{Y}^*$.
6. Compute $\mathbf{Y}$ from $\mathbf{Y}^*$.

where $\mathbf{X}^*, \mathbf{S}^*,$ and $\mathbf{Y}^*$ are defined as follows:

$$x_i^* = \Phi^{-1}(F_i(x_i)),\ i = 1,\ldots,n$$
$$s_j^* = \Phi^{-1}(F_j(s_j)),\ j = 1,\ldots,m,\ \text{and}$$
$$y_k^* = \Phi^{-1}(F_k(y_k)),\ k = 1,..,n.$$

and F represents the cumulative distribution function of the individual marginal variables. Note that this approach only requires that the marginal distribution of the individual variables be identified. The joint distribution is then approximated by the multivariate normal copula. However, since the multivariate normal copula is only an approximation of the true joint distribution, C-GADP does not provide complete data utility. More specifically, using pair-wise rank order correlations are adequate to preserve all monotonic (both linear and non-linear) relationships. However, if the data consists of

non-monotonic relationships (∪-shaped or ∩-shaped or sine wave), then C-GADP perturbed variables will not maintain such relationships. While this represents one problem with C-GADP, considering the ability of other approaches for perturbing non-normal variables, C-GADP represents a significant improvement. Sarathy et al. (2002) also provide an example application of the C-GADP approach. Interested readers can also visit the following web site for the example data set that was used in the study.

http://gatton.uky.edu/faculty/muralidhar/maskingpapers/

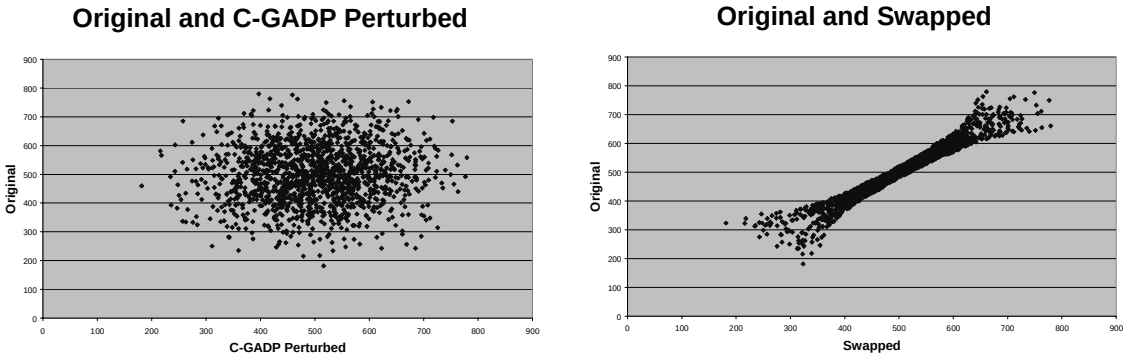
Thus, in terms of data utility, C-GADP is capable of:

(1) Maintaining the marginal distribution of the perturbed variables to be the same as that of the (original) confidential variables, even if the original variables have marginal distributions that are not normal, and
(2) Maintaining monotonic dependence (linear or non-linear) between variables that can be captured through pair-wise rank order correlation.
(3) In terms of disclosure risk, like GADP, C-GADP also assures that, given **S**, **X** and **Y** are independent. Hence, C-GADP perturbed variables do not provide an intruder with no additional information.

Enhancing and extending the copula-based approach for perturbation represents an important direction for future research.

The Data Shuffle

One of the complaints against all "perturbation" approaches is that the perturbed values are "not the same" as the original values. In other words, the process of masking "changes" the original values. Although from a statistical perspective this makes little difference, it appears to make a big difference from a practical perspective. The Wall Street Journal (February 14, 2001) quotes a Census Bureau researcher as stating, "… users have found this extremely irritating and unacceptable". One technique that facilitates releasing microdata without "modifying" the original values is data swapping. For numerical variables, Moore (1996) suggests the use of the "rank-based proximity swap" and provides valuable insights into the swapping process. However, compared to perturbation methods, swapping provides lower data utility and perhaps, more importantly, far higher risk of disclosure. For illustration purposes, consider the following figures that provide the original and C-GADP perturbed values and the original and swapped values for a particular data set.



The figure shows practically no relationship between the original and C-GADP perturbed values. By contrast, there is a very strong relationship between the original and swapped values, especially in the middle sections of the data compared to the ends (and hence the "bow tie" effect). Even a lay person would conclude from the above figures that it is far more likely that an intruder would be able to predict the value of the original variable with far higher accuracy using the swapped values than the C-GADP perturbed values. It is also likely that, because of such a relationship, swapped values also result in far higher risk of identity disclosure. We have actually verified that when swapping is used, for the data set illustrated above, it is possible to re-identify 1035 of the 1500 observations or approximately 69% of the observations. By contrast, when C-GADP is used, only 2 of the 1500 observations (or 0.13%) are re-identified. Further, swapping also results in modifying relationships between variables (Moore 1996). Finally, unlike other masking procedures, there is very little theoretical support for swapping. Although Moore (1996) has derived some results with respect to swapping, a strong theoretical basis that exists for perturbation methods does not exist. Thus, other than *user acceptance*, there is very little in favor of swapping. However, user acceptance is an important consideration in microdata release.

Thus, there is a need to develop a new technique that is capable of combining the benefits of both perturbation and swapping. Such an approach must provide the same data utility and disclosure risk characteristics as GADP or C-GADP, but must use only the original (unmodified) data in this procedure. We have developed such a method that we refer to as "The Data Shuffle". Data shuffling was developed on the same strong theoretical foundations of GADP and C-GADP and hence possesses the same data utility and disclosure risk characteristics as the perturbation techniques. However, unlike the perturbation methods, it does not modify the original values and the original values are directly used in masking the data. Data shuffling and data swapping differ in one important respect that unlike data swapping, data shuffling does not exchange values between records. Values are "truly" shuffled and consequently, the value of the i^{th} record could possibly be assigned to the j^{th} record, that of the j^{th} record to the k^{th} record, etc. Thus, the actual values in the shuffled data are indeed the original values, just shuffled in such a manner as to:

(1) Maintain the marginal distribution of the individual variables exactly (benefit of swapping),
(2) Maintain all monotonic relationships between confidential variables to be the same before and after shuffling (benefit of perturbation),
(3) Maintain all monotonic relationships between confidential and non-confidential variables to be the same before and after shuffling (benefit of perturbation), and
(4) Maintain disclosure risk to a minimum (benefit of perturbation).

We are currently in the process of developing a manuscript that describes the theoretical foundation for data shuffling, provides an illustration of data shuffling (both for simulated and real data), evaluates data utility and disclosure risk characteristics, compares its performance to other masking techniques, and addresses implementation issues relating to small data sets. Interested readers are requested to contact either author by email (sarathy@okstate.edu or krishm@uky.edu) if they would like to request a copy of the manuscript when it is completed.

REFERENCES

Dalenius, T. 1977. Towards a Methodology for Statistical Disclosure Control. *Statistisktidskrift* **5** 429-444.

Duncan, G.T. and D. Lambert 1986. Disclosure-Limited Data Dissemination. *Journal of the American Statistical Association* **81** 10-18.

Fienberg, S.E., U.E. Makov, and A.P. Sanil 1997. A Bayesian Approach to Data Disclosure: Optimal Intruder Behavior for Continuous Data. *Journal of Official Statistics* **13** 75-89.

Fuller, W.A. 1993. Masking Procedures for Microdata Disclosure Limitation. *Journal of Official Statistics* **9** 383-406.

Kim, J. 1986. A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation. *Proceedings of the American Statistical Association, Survey Research Methods Section*, ASA, Washington D.C. 370-374.

Kruskal, W.H. 1958. Ordinal Measures of Association. *Journal of the American Statistical Association* **53** 814-861.

Moore, R. 1996. Controlled Data Swapping Techniques for Masking Public Use Data Sets. U.S. Bureau of the Census, Statistical Research Division Report (rr96/04).

Muralidhar, K., R. Parsa, and R. Sarathy 1999. A General Additive Data Perturbation Method for Database Security. *Management Science* **45** 1399-1415.

Muralidhar, K., R. Sarathy, R., and R. Parsa 2001. An Improved Security Specification for Data Perturbation with Implications for E-Commerce. *Decision Sciences* **32** 683-698.

Muralidhar, K. and R. Sarathy 2003. A Theoretical Basis for Perturbation Methods. *Statistics and Computing* (forthcoming).

Sarathy, R., K. Muralidhar, and R. Parsa 2002. Perturbing Non-Normal Confidential Attributes: The Copula Approach. *Management Science* **48** 1613-1627.

Sullivan, G. 1989. The Use of Added Error to Avoid Disclosure in Microdata Releases. *Unpublished Ph.D. Dissertation*, Iowa State University, Ames, Iowa.

Sullivan, G. and W.A. Fuller. 1989. The Use of Measurement Error to Avoid Disclosure. *Proceedings of the American Statistical Association, Survey Research Methods Section*, 802-807.

Tendick, P. 1991. Optimal Noise Addition for Preserving Confidentiality in Multivariate Data. *Journal of Statistical Planning and Inference* **27** 341-353.

Tendick, P. 1992. Assessing the Effectiveness of Noise Addition, Method of Preserving Confidentiality in the Multivariate Normal Case. *Journal of Statistical Planning and Inference* **31** 273-282.

Tendick, P. and N. Matloff 1994. A Modified Random Perturbation Method for Database Security. *ACM Transactions on Database Systems*, **19** 47-63.

Traub, J.F., Y. Yemini, and H. Wozniakowski, 1984. The Statistical Security of a Statistical Database. *ACM Transactions on Database Systems* **9** 672-679.

Willenborg, L. and T. De Waal 2001. *Elements of Statistical Disclosure Control*. Springer, New York.

Yancey, W.E., Winkler, W.E., and Creecy, R. H. 2002. Disclosure Risk Assessment in Perturbative Microdata Protection, in (J. Domingo-Ferrer, ed.) *Inference Control in Statistical Databases*, Springer: New York.