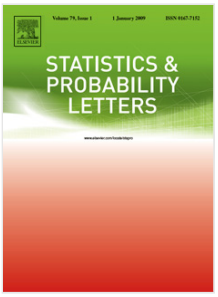
Maintaining tail dependence in data shuffling using t copula

Mario Trottini, Krish Muralidhar, Rathindra Sarathy

Accepted Manuscript



Maintaining tail dependence in data shuffling using t copula

Mario Trottini, Krish Muralidhar, Rathindra Sarathy

PII: S0167-7152(10)00342-1
DOI: 10.1016/j.spl.2010.12.002
Reference: STAPRO 5858

To appear in: *Statistics and Probability Letters*

Received date: 2 June 2010
Revised date: 24 November 2010
Accepted date: 6 December 2010

Please cite this article as: Trottini, M., Muralidhar, K., Sarathy, R., Maintaining tail dependence in data shuffling using t copula. *Statistics and Probability Letters* (2010), doi:10.1016/j.spl.2010.12.002

Maintaining Tail Dependence in Data Shuffling using t Copula

Mario Trottini [a,1], Krish Muralidhar [b], Rathindra Sarathy [c]

[a] *University of Alicante, P.O. Box 99, Alicante, Spain*

[b] *University of Kentucky, Lexington KY 40506*

[c] *Oklahoma State University, Stillwater OK 74078*

Abstract

Data shuffling is a recently proposed technique for masking numerical data where the confidential values are shuffled between records while maintaining all monotonic relationships between the variables in the data set. Data shuffling is based on the multivariate normal copula which assumes that there is no tail dependence in the data set. In many practical situations, however, tail dependence plays a crucial role in decision making. Hence, it is desirable that the data masking procedure be capable of preserving tail dependence when present. In this study, we provide a new data shuffling approach based on t copulas that is capable of maintaining tail dependence in the masked data in a large number of applications.

Keywords: Statistical confidentiality, copulas, data shuffling, disclosure risk, data dissemination, tail dependence.

1. Introduction

Many organizations (private, public, and governmental) gather, store, analyze, share, and disseminate large quantities of data. Often some of the data that has been gathered by these organizations are considered sensitive. When this is the case statistical disclosure limitation techniques (*data masking*) are applied to the collected data to produce a new data set that ideally should be *safe* from attack of potential intruders and *useful* for the statistical analysis that legitimate users might want to perform (for a review of masking techniques see for example Muralidhar, and Sarathy 2003).

Developments in data masking techniques have been driven by the desire to provide users with masked data that are capable of maintaining the same statistical characteristics as the original data. Early techniques relied on adding noise to the confidential data to mask the original values (Traub et al. 1984). However, the original noise addition techniques resulted in modifying the marginal distribution and relationships among variables. This led to the development of modified noise addition techniques that attempted to maintain linear relationships between the confidential variables (Kim 1986). Subsequent developments led to masking techniques based on linear models that were capable of maintaining linear relationships between all variables such as multiple imputation (Rubin 1993), general additive data perturbation (Muralidhar et al. 1999), and information preserving statistical obfuscation (Burridge 2003). The linear models were not capable of maintaining the marginal distribution and non-linear relationships. This led to the development of copula based perturbation (Sarathy et al. 2002, Trottini

[1] Corresponding address: Dpto. de Estadística e I.O., Universidad de Alicante, Apartado de Correos 99, 03080, Alicante, Spain, FAX: +34-965903667, Tel.: +34-965903531.
E-mail addresses: mario.trottini@ua.es, (M. Trottini), krish.muralidhar@uky.edu (K. Muralidhar), rathin.sarathy@okstate.edu (R. Sarathy)

et al. 2008) and skew t perturbation (Lee et al. 2010) that allowed a flexible modeling of the marginal distributions and multivariate dependence. The copula based approach provides masked data that maintain all monotonic relationships between variables but preserves marginals only asymptotically when the marginal distributions are known prior to the masking. Skew t perturbation, on the other hand, relies on a parametric family much more flexible than the multivariate normal but still preserves linear and non-linear relationships and marginals only when the multivariate skew t distribution is a suitable model for the original data (and only asymptotically for the marginals).

Recently, Muralidhar and Sarathy (2006) proposed a Data Shuffling procedure for masking sensitive numerical data (from now on *G-DS procedure*) that enhanced the copula based perturbation approach. In G-DS, the original sensitive data values are "shuffled" among the records so as to preserve the marginal distribution of the sensitive variables as well as monotonic relationships between all (both sensitive and non-sensitive) variables. In terms of security, since the shuffled data are generated as a function only of the values of the original non-sensitive variables, the masked data do not provide any additional information to the intruder. For a comprehensive discussion of G-DS, please refer to Muralidhar and Sarathy (2006).

In keeping with this trend, in this paper, we propose an extension to the G-DS method that is capable of preserving tail dependence in addition to the other benefits derived from the original G-DS approach. The paper is organized as follows. In the next section, we briefly describe the G-DS procedure. Section 3 introduces the notion of tail dependence and explains limitation of G-DS in dealing with it. In section 4 an alternative approach based on t-copula is presented. Advantages and limitations of the new t-copula based data shuffling procedure compared to the standard G-DS are discussed in section 5. Finally in section 6 we summarize the main results of the paper and outline ideas of future work.

2. The G-DS procedure

Consider a data set comprising a set of numerical confidential data **X** of dimension M (variables) and N (records) and a set of numerical nonconfidential data **S** of dimension L and N (the nonconfidential variables in **S** are for the same records as in **X**) [2]. We assume that: (i) the empirical cumulative distribution function for the *j*th variable in **X** can be well approximated by a strictly increasing cumulative distribution function (cdf) F_j (*j=1, ..., M*); (ii) the empirical distribution function for the *k*th variable in **S** can be well approximated by a strictly increasing cdf G_k (*k= 1, ..., L*); and (iii) the empirical joint cumulative distribution function for the (M+L) variables in the data can be well approximated by a multivariate continuous cdf $F_{\mathbf{X},\mathbf{S}}$. Let $x_{i,j}$ and $x_{(i),j}$ represent the *i*th (unordered) and the *i*th ordered observation for the *j*th confidential variable (i.e $rank(x_{(i),j})=i$) and let $\mathbf{x}_i$ and $\mathbf{s}_i$ represent $(1\times M)$ and $(1\times L)$ single-observation vectors from **X** and **S** respectively. The data shuffling procedure of Muralidhar

[2] In order to simplify notation, in the rest of the paper we use **X** (**S**) to denote both the confidential (nonconfidential) data and the set of confidential (nonconfidential) variables. In each case, the correct interpretation of **X** (**S**) should be clear from the context.

and Sarathy (from now on *DS procedure*) can be described as follows (see Muralidhar and Sarathy 2006 page 661):

- *Step 1*. For *i=1,…,N,* generate *perturbed* vectors $\mathbf{y}_i^p$ from the conditional distribution of $\mathbf{X}$ given $\mathbf{S}=\mathbf{s}_i$. Let $\mathbf{Y}^p$ be the corresponding $(N\times M)$ matrix of perturbed values and let $y^p_{(i),j}$ represent the *i*th ordered observation of the *j*th perturbed variable (*j*th column of $\mathbf{Y}^p$).
- *Step 2*. For *i=1,…,N,* and *j=1,…,M* replace $y^p_{(i),j}$ with $x_{(i),j}$. Let $\mathbf{Y}$ represent the corresponding $(N\times M)$ matrix of *shuffled* values.
- *Step 3*. Release the reordered (or shuffled) data set (**S**, **Y**).

Table 1: ***DS - procedure***

In the DS procedure perturbed values for the sensitive variables are generated according to the conditional distribution of **X** given **S** (step 1). The perturbed values are then used to make a "smart" shuffling of the original sensitive variables that are finally released (steps 2 and 3). The shuffling is "smart" since it is made according to the ranks of values for the sensitive variables generated from the conditional distribution of **X** given **S**. As such the shuffling preserve not only the marginal distributions of (**X**,**S**) but also the joint distribution of the original data achieving the maximum data utility.

Unfortunately in many real applications, the conditional distribution of **X** given **S** cannot be derived and the DS procedure, as described in table 1 cannot be implemented. The heuristic solution proposed by Muralidhar and Sarathy (2006), that we will refer to as the *G-DS procedure*, consists of generating the perturbed values for the sensitive variables in step 1 from a conditional distribution which is not the "true" conditional of **X** given **S** but is obtained from the joint distribution of a random vector that has the same univariate margins and the same Kendall's tau (or Spearman's rho) correlation matrix of the "true" joint distribution for (**X, S).** Thus, to the extent to which Kendall's tau (or Spearman's rho) is an appropriate measure of dependence for the original data, the G-DS procedure provides a shuffled data set that preserves both the marginals and the relevant features of dependence structure of the original data.

Since both Kendall's tau and Spearman's rho are copula based measures of dependence, i.e. they only depend on the copula underlying the joint distribution of (**X**,**S**), regardless of the margins (see Joe, 1997 p.32), copulas provide a natural tool for the required implementation. Assuming a Gaussian copula model for (**X**,**S**), the cdf of (**X**,**S**) can be represented as follows:

$$F_{\mathbf{X},\mathbf{S}}(x_1,...,x_M,s_1,...,s_M)=\Phi^{\rho}\left(\phi^{-1}(F_1(x_1)),...,\phi^{-1}(F_M(x_M)),\phi^{-1}(G_1(s_1)),...,\phi^{-1}(G_L(s_L))\right) \tag{1}$$

where Φ^{ρ} represents the cdf of a (M+L)-variate normal distribution with mean **0**, and product moment correlation matrix ρ and ϕ^{-1} is the quantile function of a univariate standard normal distribution. The copula parameter ρ, is related to the Spearman's rho and Kendall's tau correlation matrix of (**X**,**S**), that we denote respectively by $\rho_S=\{\rho^S_{i,j}\}_{i,j=1,...M+L}$, and $\rho_t=\{\rho^t_{i,j}\}_{i,j=1,...M+L}$ by the formulas:

$$\rho_{i,j}=2\cdot\sin\left(\pi\cdot\rho^S_{i,j}/6\right),\quad i,j=1,...M+L; \tag{2}$$

$$\rho_{i,j} = \sin\left(\pi \cdot \rho^{\tau}_{i,j} / 2\right), \qquad i, j = 1,...M + L. \tag{3}$$

Thus $\boldsymbol{\rho}$ can be estimated evaluating either the Spearman's or the Kendall's tau correlation matrix of the original data and then using (2) or (3). Denoting by $\hat{\boldsymbol{\rho}}$ the estimate of $\boldsymbol{\rho}$, the estimated copula model becomes:

$$\hat{F}_{\mathbf{X},\mathbf{S}}(x_1,...,x_M,s_1,...,s_M) = \Phi^{\hat{\boldsymbol{\rho}}}\left(\phi^{-1}(F_1(x_1)),...,\phi^{-1}(F_M(x_M)),\phi^{-1}(G_1(s_1)),...,\phi^{-1}(G_L(s_L))\right) \tag{4}$$

Note that, regardless of the true distribution of (**X,S**), by construction the model in (4) exactly preserves the univariate margins and the Kendall's tau (or Spearman's rho) correlation matrix of the "true" distribution (depending on whether (2) or (3) is used for estimating $\boldsymbol{\rho}$). Let $\mathbf{X}^*$ and $\mathbf{S}^*$ be random vectors defined as follows,

$$\begin{aligned} &\mathbf{X}^* = \left(X_1^*,...,X_M^*\right); \qquad \mathbf{S}^* = \left(S_1^*,...,S_L^*\right); \\ &X_j^* = \phi^{-1}\left(F_j(X_j)\right), \quad j=1,...,M; \qquad S_k^* = \phi^{-1}\left(G_k(S_k)\right), \quad k=1,...,L. \end{aligned} \tag{5}$$

Under (4) $(\mathbf{X}^*,\mathbf{S}^*)$ follows a multivariate standard normal distribution with covariance (and correlation) matrix $\hat{\boldsymbol{\rho}}$. From basic properties of the multivariate normal distribution the conditional distribution of $\mathbf{X}^*|\mathbf{S}^*=\mathbf{s}^*$ is also multivariate normal ,

$$\mathbf{X}^*|\mathbf{S}^* = \mathbf{s}^* \sim MN\left(\boldsymbol{\mu}_{\mathbf{X}^*.\mathbf{S}^*}, \boldsymbol{\Sigma}_{\mathbf{X}^*\mathbf{X}^*.\mathbf{S}^*}\right) \tag{6}$$

with vector mean $\boldsymbol{\mu}_{\mathbf{X}^*.\mathbf{S}^*}$ and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{X}^*\mathbf{X}^*.\mathbf{S}^*}$ given by:

$$\boldsymbol{\mu}_{\mathbf{X}^*.\mathbf{S}^*} = \hat{\boldsymbol{\rho}}_{\mathbf{X}^*\mathbf{S}^*}\left(\hat{\boldsymbol{\rho}}_{\mathbf{S}^*\mathbf{S}^*}\right)^{-1}\mathbf{s}^*; \qquad \boldsymbol{\Sigma}_{\mathbf{X}^*\mathbf{X}^*.\mathbf{S}^*} = \hat{\boldsymbol{\rho}}_{\mathbf{X}^*\mathbf{X}^*} - \hat{\boldsymbol{\rho}}_{\mathbf{X}^*\mathbf{S}^*}\left(\hat{\boldsymbol{\rho}}_{\mathbf{S}^*\mathbf{S}^*}\right)^{-1}\hat{\boldsymbol{\rho}}_{\mathbf{S}^*\mathbf{X}^*} \tag{7}$$

where in (7) we used the partition

$$\hat{\boldsymbol{\rho}} = \begin{pmatrix} \hat{\boldsymbol{\rho}}_{\mathbf{X}^*\mathbf{X}^*} & \hat{\boldsymbol{\rho}}_{\mathbf{X}^*\mathbf{S}^*} \\ \hat{\boldsymbol{\rho}}_{\mathbf{S}^*\mathbf{X}^*} & \hat{\boldsymbol{\rho}}_{\mathbf{S}^*\mathbf{S}^*} \end{pmatrix} \text{ with dimension } \begin{pmatrix} M\times M & M\times L \\ L\times M & L\times L \end{pmatrix}. \tag{8}$$

The G-DS procedure can be then described as follows:

- Step *1.a* For each $i = 1,\ldots,N$ generate $\mathbf{y}_i^* = \left(y_{i,1}^*,\ldots,y_{i,M}^*\right)$ from the conditional distribution of $\mathbf{X}^*|\mathbf{S}^*=\mathbf{s}_i^*$ in (6) to result in $\mathbf{Y}^*$.
- *Step 1.b*: For each $i = 1,\ldots,N$ and for each $j=1,\ldots,M$ perform the reverse mapping, $y_{i,j}^p = \phi\left(F_{X_j}^{-1}(y_{i,j}^*)\right)$ to result in $\mathbf{Y}^p$.
- *Steps 2 and 3*: the same as in the DS procedure in table 1.

Table 2: ***G-DS-procedure***

As observed by Muralidhar and Sarathy (2006) the algorithm in table 2 can be generalized to the case in which the marginal distributions of (**X**,**S**) are unknown. First of all, since the data shuffling procedure only uses the ranks of the perturbed values, *Step 1.b* in table 2 is, in fact, unnecessary (and thus the knowledge of the marginal distribution for the sensitive variables is not longer required) . Under the assumption of strictly increasing marginal cdf for the *j*th sensitive variables X_j, $rank(y_{i,j}^p) = rank(y_{i,j}^*)$, so that in the above procedure in *Step 1.b* we could set

directly $\mathbf{Y}^{p} = \mathbf{Y}^{*}$. In addition if the marginal distributions of the nonconfidential variables are unknown, Muralidhar and Sarathy (2006) suggest to replace the matrix of the transformed nonconfidential variables $\mathbf{S}^{*}$ in (5) with an estimate $\hat{\mathbf{S}}^{*}$ by approximating the (*i, k*) element of $\mathbf{S}^{*}$ with:

$$\hat{s}^{*}_{i,k} = \phi^{-1}\left(\left[(i) - 0.5\right]/n\right), \qquad i = 1,...,N; \quad k = 1,...,L \tag{9}$$

where (*i*) represents the rank order of $s_{i,k}$, and $\hat{s}^{*}_{i,k}$ the (*i,k*) element of $\hat{\mathbf{S}}^{*}$.

The G-DS algorithm in table 2 (with the simplification and generalizations that we discussed above) guarantees that although the shuffling of the original data is done according to a joint distribution possibly different from the true joint distribution of (**X**,**S**), the shuffling preserves important features of such distribution, namely the marginal distributions and the dependence structure (as measured by Spearman's rho or Kendall's tau correlation matrix of the original data). This might seem to be a reasonable objective for most practitioners. Quoting Joe (1997, p.16),

> *"My view of multivariate modeling, based on experience with multivariate data is that* ***models should try to capture*** *important characteristics, such as* ***density shapes for univariate margins*** *and* ***the appropriate dependence structure*** *and otherwise be simple as possible".*

In many applications, however, the Spearman's rho or Kendall's tau correlation matrix, constitute only a specific aspect and a partial representation of the dependence structure of the original data (both are measures of "concordance", see, for example, Joe 1997, chapter 2). The nature of dependence can take a variety of other forms that the Gaussian copula model is not able to recover and that might be of great importance in applications. One of such notion of dependence is *tail dependence* that we discuss next.

3. Tail Dependence

The following definition formalizes the notion of *lower* and *upper tail dependence* (see Nelsen 2006, p. 214). Let *X* and *Y* be continuous random variables with distributions functions *F* and *G* respectively. The *upper tail dependence parameter* λ_U is the limit (if it exists) of the conditional probability that *Y* is greater than the 100α percentile of *G* given that *X* is greater than the 100α percentile of *F* as α approaches 1, i.e.,

$$\lambda_U = \lim_{\alpha \to 1^-} P(Y > G^{-1}(\alpha) | X > F^{-1}(\alpha)). \tag{10}$$

Similarly the *lower tail dependence parameter* λ_L is defined as

$$\lambda_L = \lim_{\alpha \to 0^+} P(Y \leq G^{-1}(\alpha) | X \leq F^{-1}(\alpha)). \tag{11}$$

Positive values of λ_L (λ_U) indicate that the joint distribution of (X,Y) tends to generate joint extreme events in the lower (upper) tails of the marginal distributions of X and Y. The concept of tail dependence has important implications in many applications and has been shown to be prevalent in economic and actuarial data (see for example, Demarta and McNeil 2004 and Frees and Valdez 1998).

It is important to notice that the measures of tail dependence in (10) and (11) are copula based, i.e. they only depend on the copula that defines the joint distribution

of the data regardless of the margins (see Joe, 1997 p. 33). Not surprisingly copulas have been often used to model data in the presence of tail dependence. Caillault and Guegan (2005), for example, show that the daily closing level at the Thai SET index, Malaysian KLCI index, and the Indonesian JCI index exhibit tail dependence which is modeled using t copulas. Patton (2006) shows similar results for daily Deutsche mark to U.S. dollar and Japanese Yen to U.S. dollar exchange rates. Frees and Valdez (1998) discuss the use of copulas for modeling tail dependence for insurance company data on losses and expenses. For imperfect correlated variables the Gaussian Copula implies tail independence (i.e. $\lambda_L=\lambda_U=0$). Thus, any Gaussian copula model, as the model in (4), implicitly assumes that no tail dependence is present in the data. This suggests that the G-DS procedure of Muralidhar and Sarathy (2006) should be expected to perform poorly and an alternative approach should be used in the presence of tail dependence. In the next section we present such an alternative based on the t-copula.

4. Data Shuffling using t Copula

In this section, we describe the use of the t copula for performing data shuffling to maintain tail dependence. Assuming a t-copula model for (**X**,**S**) , the cdf of (**X**,**S**) can be represented as follows:

$$F_{\mathbf{X},\mathbf{S}}(x_1,...,x_M,s_1,...,s_M) = t_{\boldsymbol{\rho},\upsilon}\left(t_\upsilon^{-1}(F_1(x_1)),...,t_\upsilon^{-1}(F_M(x_M)),t_\upsilon^{-1}(G_1(s_1)),...,t_\upsilon^{-1}(G_L(s_L))\right) \tag{12}$$

where $t_{\boldsymbol{\rho},\upsilon}$ represents the joint cdf of a k-variate Student t distribution with mean **0**, product moment correlation matrix $\boldsymbol{\rho}$, and ν degrees of freedom and t_ν^{-1} is the quantile function of a univariate t-distribution with ν degrees of freedom. The Gaussian-copula model in (1), that was used in the G-DS, procedure can be obtained as limiting of the t copula model in (12) as $\upsilon \to \infty$ and shares many common characteristics with the t copula model. However, they differ on one important characteristic, namely, that while the multivariate normal copula model is not able to capture the phenomenon of dependence in extreme values, the t copula model provides this important ability through the use of the parameter υ. For the t-copula model in (12), in fact, the lower and upper tail dependence coefficients (λ_L and λ_U) for an arbitrary pair of random variables in (**X**,**S**) are given by (see Demarta and McNeil, 2005, p.114):

$$\lambda_L = \lambda_U = 2\cdot t_{\nu+1}\left(-\sqrt{\upsilon+1}\cdot\sqrt{1-\rho}/\sqrt{1+\rho}\right), \tag{13}$$

where ρ is the element of $\boldsymbol{\rho}$ corresponding to the particular pair of random variables considered . As shown in (13), for a given υ, tail dependence increases with ρ; and for a given ρ tail dependence decreases with υ (i.e. smaller values of υ indicate higher tail dependence). As for the Gaussian copula, the scale parameter $\boldsymbol{\rho}$ of the t-copula is related to the Kendall's tau correlation matrix of (**X**,**S**) through formula (3). Thus, also in this case, $\boldsymbol{\rho}$ can be estimated evaluating the Kendall's tau correlation matrix of the original data and then using (3). The remaining parameter υ can be estimated by maximum likelihood with the matrix $\boldsymbol{\rho}$ held fixed (see Demarta and McNeil 2005, section 4.2). Denoting by $\hat{\boldsymbol{\rho}}$ and $\hat{\upsilon}$ such estimates the estimated copula model becomes:

$$\hat{F}_{\mathbf{X},\mathbf{S}}(x_1,...,x_M,s_1,...,s_M)=t_{\hat{\boldsymbol{\rho}},\hat{\upsilon}}\left(t_{\hat{\upsilon}}^{-1}(F_1(x_1)),...,t_{\hat{\upsilon}}^{-1}(F_M(x_M)),t_{\hat{\upsilon}}^{-1}(G_1(s_1)),...,t_{\hat{\upsilon}}^{-1}(G_L(s_L))\right) \tag{14}$$

Let $\mathbf{X}^*$ and $\mathbf{S}^*$ be random vectors defined as follows,

$$\begin{aligned}&\mathbf{X}^*=\left(X_1^*,...,X_M^*\right); \qquad \mathbf{S}^*=\left(S_1^*,...,S_L^*\right);\\ &X_j^*=t_{\hat{\upsilon}}^{-1}\left(F_j(X_j)\right), \qquad j=1,...,M; \qquad S_k^*=t_{\hat{\upsilon}}^{-1}\left(G_k(S_k)\right), \qquad k=1,...,L.\end{aligned} \tag{15}$$

Under (14) the distribution of $(\mathbf{X}^*,\mathbf{S}^*)$ is a multivariate Student's t with mean vector $\mathbf{0}$, scale matrix $\hat{\boldsymbol{\rho}}$, and $\hat{\upsilon}$ degrees of freedom. In addition, the conditional distribution of $\mathbf{X}^*$ given $\mathbf{S}^*=\mathbf{s}^*$ is also multivariate Student's t with location parameter $\boldsymbol{\mu}_{\mathbf{X}^*.\mathbf{s}^*}$, scale matrix $\boldsymbol{\Sigma}_{\mathbf{X}^*\mathbf{X}^*.\mathbf{S}^*}$ and $\upsilon_{\mathbf{X}^*.\mathbf{S}^*}$ degrees of freedom,

$$\mathbf{X}^*|\mathbf{S}^*=\mathbf{s}^*\sim t\left(\boldsymbol{\mu}_{\mathbf{X}^*.\mathbf{s}^*},\boldsymbol{\Sigma}_{\mathbf{X}^*\mathbf{X}^*.\mathbf{S}^*},\upsilon_{\mathbf{X}^*.\mathbf{S}^*}\right) \tag{16}$$

where (see, for example, Arslan, 2004, proposition 4):

$$\boldsymbol{\mu}_{\mathbf{X}^*.\mathbf{S}^*}=\hat{\boldsymbol{\rho}}_{\mathbf{X}^*\mathbf{S}^*}\left(\hat{\boldsymbol{\rho}}_{\mathbf{S}^*\mathbf{S}^*}\right)^{-1}\mathbf{s}^*; \qquad \upsilon_{\mathbf{X}^*.\mathbf{S}^*}=\hat{\upsilon}+L; \tag{17}$$

$$\boldsymbol{\Sigma}_{\mathbf{X}^*\mathbf{X}^*.\mathbf{S}^*}=\left(\hat{\upsilon}+L\right)^{-1}\left[\hat{\upsilon}+\mathbf{s}^{*T}\left(\hat{\boldsymbol{\rho}}_{\mathbf{S}^*\mathbf{S}^*}\right)^{-1}\mathbf{s}^*\right]\cdot\left[\hat{\boldsymbol{\rho}}_{\mathbf{X}^*\mathbf{X}^*}-\hat{\boldsymbol{\rho}}_{\mathbf{X}^*\mathbf{S}^*}\left(\hat{\boldsymbol{\rho}}_{\mathbf{S}^*\mathbf{S}^*}\right)^{-1}\hat{\boldsymbol{\rho}}_{\mathbf{S}^*\mathbf{X}^*}\right] \tag{18}$$

and in (18) we used the same partition of $\hat{\boldsymbol{\rho}}$ as in (8). The heuristic implementation of the DS procedure that we propose (and that we will refer to as *t-DS procedure*) can be then described as follows:

- Step *1.a* For each $i=1,...,N$ generate $\mathbf{y}_i^*=\left(y_{i,1}^*,\ldots,y_{i,M}^*\right)$ from the conditional distribution of $\mathbf{X}^*|\mathbf{S}^*=\mathbf{s}_i^*$ in (16) to result in $\mathbf{Y}^*$.
- *Step 1.b*: For each $i=1,...,N$ and for each $j=1,...,M$ perform the reverse mapping, $y_{i,j}^p=t_\upsilon\left(F_{X_j}^{-1}(y_{i,j}^*)\right)$ to result in $\mathbf{Y}^p$.
- *Steps 2 and 3*: the same as in the DS procedure in table 1.

Table 3 : ***t-DS-procedure***

Following Muralidhar and Sarathy (2006) the algorithm in table 3 can be generalized to the case in which the marginal distributions of (**X**,**S**) are unknown by setting $\mathbf{Y}^p=\mathbf{Y}^*$ (thus eliminating *Step 1.b*) and by replacing the matrix of the transformed nonconfidential variables $\mathbf{S}^*$ in (15) with an estimate $\hat{\mathbf{S}}^*$ by approximating the (*i, k*) element of $\mathbf{S}^*$ with:

$$\hat{s}_{i,k}^*=t_{\hat{\upsilon}}^{-1}\left(\left[(i)-0.5\right]/n\right), \qquad i=1,...,N; \quad k=1,...,L \tag{19}$$

where *(i)* represents the rank order of $s_{i,k}$, and $\hat{s}_{i,k}^*$ the (*i,k*) element of $\hat{\mathbf{S}}^*$.

The t-DS procedure discussed above can be thought as an extension of the G-DS procedure discussed by Muralidhar and Sarathy (2006). The two, in fact, coincide if no tail dependence is present. Both procedures preserve the marginal distributions of the original data exactly, and the Kendall's tau correlation matrix of the original data, asymptotically (as the number of observations tends to infinity). The data shuffled using the t copula model provides the additional advantage that it allows to model tail dependence as well. In terms of privacy protection, the proposed t-DS procedure relies on the same definition of disclosure, and the same disclosure scenario as the original G-DS procedure and

provides the same privacy protection as the original G-DS procedure (for details see Muralidhar and Sarathy 2006 and the references therein). This disclosure risk scenario assures that, given **S**, **X** and **Y** are independent. Hence, releasing the masked microdata in place of the original data does not result in increased risk of disclosure over and above the information available from **S** and summary information.

5. An Empirical Comparison of Multivariate Normal and t copula Shuffling

In this section, we present the results of a simulation study to compare the performance of the proposed t-DS procedure with the performance of the standard G-DS procedure discussed by Muralidhar and Sarathy (2006). In the simulation study we consider a bivariate random vector **V**=(X,S) with standard normal marginals. It is assumed that variable X is confidential and variable S is not confidential. For the random vector **V**, twelve distributions are considered. The first six refer to a bivariate t-copula model as in (12) with different values of the t-copula parameters ρ and υ to allow for different degrees of concordance (as measured as Kendall's tau) and tail dependence. The remaining six distributions refer to a bivariate random variable with standard normal margins and a mixture copula $C_{mix}^{\rho,\vartheta,w}$ given by:

$$C_{mix}^{\rho,\vartheta,w}(u_1,u_2) = (1-w)\cdot C_{Ga}^{\rho}(u_1,u_2) + w\cdot C_{A}^{\vartheta}(u_1,u_2) \tag{20}$$

where $(u_1,u_2) \in [\mathbf{0},\mathbf{1}]^2$, $0<w<1$; C_{Ga}^{ρ} is a Gaussian copula with correlation parameter ρ (see Joe, 1997 p.140); and C_{A}^{ϑ} is the (4.2.12)-Archimedean Copula discussed in Nelsen 2006 (p. 116),

$$C_{A}^{\vartheta}(u_1,u_2) = 1/\left\{1+\left[\left(u_1^{-1}-1\right)^{\vartheta}+\left(u_2^{-1}-1\right)^{\vartheta}\right]^{1/\vartheta}\right\}. \tag{21}$$

The resulting joint CDF for (X,S) is,

$$F_{\mathrm{X,S}}(x,s) = C_{mix}^{\rho,\vartheta,w}\left(\phi(x),\phi(s)\right). \tag{22}$$

It can be shown that for the model in (22) we have (see p. 215 in Nelsen 2006): $\lambda_L = w\cdot 2^{-1/\vartheta}$; $\lambda_U = w\cdot\left(2-2^{1/\vartheta}\right)$. The reason for considering the mixture copula model in (22) is that it allows to understand how well the proposed t-DS method perform compared to the G-DS method when the copula underlying the actual data distribution is not the t copula under which the t-DS procedure is built. Different choices of the mixture copula parameters (ρ,θ,w) are considered to allow for different degree of concordance between X and S, asymmetry in tail dependence, and departure from the t-copula model. The parameter vectors for each of the twelve distributions in the study with the corresponding lower and upper tail dependence coefficients, and the Kendall's tau (ρ_τ) correlation are shown in table 4. For each of the twelve distributions in table 4 we generated 1000 data sets consisting of N=5000 i.i.d realizations of the underlying random vector (X,S). To each data set both the G-DS and t-DS procedures are then applied. Since tail dependence is an asymptotic concept, we compare the performance of the two procedures in terms of bias and standard deviation of bias with respect to answers to specific queries concerning joint extreme events (joint exceedance probabilities).

Distribution	Underlying Copula	Copula parameters	λ_L	λ_U	Kendall's tau[3]
1	t-copula	ρ=0.05, υ=100	$8\cdot10^{-16}$	$8\cdot10^{-16}$	0.03
2	t-copula	ρ=0.5, υ=100	$7\cdot10^{-8}$	$7\cdot10^{-8}$	0.33
3	t-copula	ρ=0.05, υ=10	0.009	0.009	0.03
4	t-copula	ρ=0.5, υ=10	0.082	0.082	0.33
5	t-copula	ρ = 0.05, υ=1	0.311	0.311	0.03
6	t-copula	ρ = 0.4, υ=2	0.339	0.339	0.26
7	$C_{mix}^{\rho,\vartheta,w}$	ρ = –0.84, ϑ=1.8, ω=0.5	0.340	0.265	0.01
8	$C_{mix}^{\rho,\vartheta,w}$	ρ = 0.4, ϑ=1.8, ω=0.5	0.340	0.265	0.43
9	$C_{mix}^{\rho,\vartheta,w}$	ρ = –0.84, ϑ=1.3, ω=0.6	0.352	0.177	0.05
10	$C_{mix}^{\rho,\vartheta,w}$	ρ = 0.4, ϑ=1.3, ω=0.6	0.352	0.177	0.39
11	$C_{mix}^{\rho,\vartheta,w}$	ρ = –0.8, ϑ=1, ω=0.7	0.35	0	0.07
12	$C_{mix}^{\rho,\vartheta,w}$	ρ = 0.7, ϑ=1, ω=0.7	0.35	0	0.38

Table 4: ***Distributions used in the simulations with the corresponding dependence measures.***

The queries considered were of the type:

$Q_{1,\alpha}$: Find the probability that both variables take values below the α quantile of their margins.

$Q_{2,\alpha}$: Find the probability that both variables take values above the(1-α) quantile of their margins.

with α = 0.005, 0.01. Tables 5 summarizes the results of the simulation study. For each of the twelve models, in table 5 we report: (i) the average joint exceedance probability (**AEP**) corresponding to the quantile queries $Q_{1,\alpha}$ and $Q_{2,\alpha}$ evaluated using the original data; (ii) the ratio between the AEP above and the corresponding average probabilities under G-DS and t-DS procedures (ratios that we denote by $\mathbf{R_{G\text{-}DS}}$ and $\mathbf{R_{t\text{-}DS}}$ respectively). A ratio of 1 indicates that the AEP of the masked data is the same as that of original data. Ratios smaller (larger) than 1 indicate over (under) estimation; (iii) the average bias (**AB**) for the joint exceedance probabilities under the G-DS and t-DS procedures; and (iv) the standard deviation of the bias ($\mathbf{SD_B}$). Since the original exceedance probabilities are very small, in order to make the comparisons more meaningful, we have provided the value of the actual probabilities × 10^6 (thus, the true probabilities are the values provided in the table multiplied by 10^{-6}). We did the same for the average bias and the bias standard deviation.

From table 5 we observe that, as expected, the t-DS procedure and G-DS procedure produce comparable results when no tail dependence is present (distributions 1 and 2). On the other hand, in the presence of tail dependence the t-DS procedure outperforms the G-DS procedure not only when the copula underlying the distribution of the original data is the t copula under which the t-DS procedure is built (distributions 3-6) but also when the model underlying the original data is different from the t-copula model but preserves, at least

[3] The value of Kendall's tau for distributions 7-12 in table 4 have been obtained as the average Kendall's tau over the 1000 simulated data sets. For distributions 7-12 the simulation was done using the simulation algorithm described in exercise 4.15, on p. 134 of Nelsen (2006).

approximately, some symmetry in the tail dependence (as it is the case for distributions 7-8 for which the difference between λ_L and λ_U is not "too large").

Distribution	Query	AEP	$R_{G\text{-}DS}$	$R_{t\text{-}DS}$	AB G-DS	AB t-DS	SD_B G-DS	SD_B t-DS
1 $\begin{pmatrix}\lambda_L=\lambda_U=8\cdot10^{-16}\\ \rho_\tau=0.03\end{pmatrix}$	$Q_{1,0.005}$	56	1,35	1,31	14	13	140	141
	$Q_{2,0.005}$	45	1,16	0,97	6	-1	130	134
	$Q_{1,0.01}$	168	1,2	1,06	28	10	242	248
	$Q_{2,0.01}$	164	1,22	1,04	30	6	247	243
2 $\begin{pmatrix}\lambda_L=\lambda_U=7\cdot10^{-8}\\ \rho_\tau=0.33\end{pmatrix}$	$Q_{1,0.005}$	531	1,08	1	38	2	392	423
	$Q_{2,0.005}$	529	1,08	1	37	-2	407	432
	$Q_{1,0.01}$	1340	1,03	0,98	39	-29	633	646
	$Q_{2,0.01}$	1369	1,07	1	95	2	648	669
3 $\begin{pmatrix}\lambda_L=\lambda_U=0.009\\ \rho_\tau=0.03\end{pmatrix}$	$Q_{1,0.005}$	187	4,85	1,04	149	7	208	265
	$Q_{2,0.005}$	183	5,14	1,03	147	6	202	254
	$Q_{1,0.01}$	467	3,32	1	327	2	347	409
	$Q_{2,0.01}$	451	3,27	1	313	0	335	399
4 $\begin{pmatrix}\lambda_L=\lambda_U=0.082\\ \rho_\tau=0.33\end{pmatrix}$	$Q_{1,0.005}$	888	1,8	1,04	394	31	448	493
	$Q_{2,0.005}$	861	1,82	0,99	389	-6	454	489
	$Q_{1,0.01}$	1987	1,57	1,02	725	39	686	699
	$Q_{2,0.01}$	1946	1,53	0,98	677	-39	688	699
5 $\begin{pmatrix}\lambda_L=\lambda_U=0.311\\ \rho_\tau=0.03\end{pmatrix}$	$Q_{1,0.005}$	1531	47,24	0,99	1498	-12	431	593
	$Q_{2,0.005}$	1512	40,65	0,99	1475	-16	438	573
	$Q_{1,0.01}$	3096	24,42	0,99	2969	-19	630	829
	$Q_{2,0.01}$	3059	22,8	0,98	2925	-47	631	832
6 $\begin{pmatrix}\lambda_L=\lambda_U=0.339\\ \rho_\tau=0.26\end{pmatrix}$	$Q_{1,0.005}$	1679	5,88	0,99	1393	-21	493	584
	$Q_{2,0.005}$	1678	6,04	0,99	1400	-9	493	600
	$Q_{1,0.01}$	3414	4,34	0,99	2628	-21	731	850
	$Q_{2,0.01}$	3406	4,37	0,99	2627	-21	724	826
7 $\begin{pmatrix}\lambda_L=0.34,\ \lambda_U=0.26\\ \rho_\tau=0.01\end{pmatrix}$	$Q_{1,0.005}$	1650	71,14	1,15	1627	214	426	575
	$Q_{2,0.005}$	1292	54,27	0,9	1268	-150	410	567
	$Q_{1,0.01}$	3365	34,26	1,17	3266	485	627	808
	$Q_{2,0.01}$	2652	28,03	0,91	2557	-255	602	815
8 $\begin{pmatrix}\lambda_L=0.34,\ \lambda_U=0.26\\ \rho_\tau=0.43\end{pmatrix}$	$Q_{1,0.005}$	1801	2,23	1,1	993	168	514	578
	$Q_{2,0.005}$	1484	1,86	0,9	688	-160	523	575
	$Q_{1,0.01}$	3794	1,9	1,11	1802	365	781	841
	$Q_{2,0.01}$	3109	1,58	0,9	1139	-331	747	835
9 $\begin{pmatrix}\lambda_L=0.35,\ \lambda_U=0.18\\ \rho_\tau=0.05\end{pmatrix}$	$Q_{1,0.005}$	1752	36,49	1,1	1704	159	447	593
	$Q_{2,0.005}$	911	21,79	0,57	869	-681	377	536
	$Q_{1,0.01}$	3540	20,44	1,11	3366	341	629	827
	$Q_{2,0.01}$	1881	10,97	0,58	1709	-1338	558	780
10 $\begin{pmatrix}\lambda_L=0.35,\ \lambda_U=0.18\\ \rho_\tau=0.39\end{pmatrix}$	$Q_{1,0.005}$	1862	2,74	1,64	1183	728	528	562
	$Q_{2,0.005}$	1020	1,54	0,88	358	-141	502	505
	$Q_{1,0.01}$	3861	2,29	1,54	2174	1362	788	790
	$Q_{2,0.01}$	2184	1,29	0,86	487	-361	711	745
11 $\begin{pmatrix}\lambda_L=0.35,\ \lambda_U=0\\ \rho_\tau=0.07\end{pmatrix}$	$Q_{1,0.005}$	1761	35,64	1,61	1711	667	447	578
	$Q_{2,0.005}$	36	0,64	0,03	-20	-1059	128	385
	$Q_{1,0.01}$	3517	18,35	1,59	3325	1304	640	777
	$Q_{2,0.01}$	144	0,76	0,07	-45	-2067	252	550
12 $\begin{pmatrix}\lambda_L=0.35,\ \lambda_U=0\\ \rho_\tau=0.38\end{pmatrix}$	$Q_{1,0.005}$	2078	3,29	2,1	1447	1089	530	563
	$Q_{2,0.005}$	384	0,6	0,39	-260	-602	395	437
	$Q_{1,0.01}$	4304	2,67	1,94	2694	2085	769	783
	$Q_{2,0.01}$	925	0,57	0,42	-688	-1297	593	643

Table 5: ***Exceedance probabilities for original, G-DS and t-DS shuffled data.***

In all these cases (distributions 1-8) the joint exceedance probability corresponding to the quantile queries $Q_{1,\alpha}$ and $Q_{2,\alpha}$ evaluated using the original data and the t-DS data are very similar (i.e. the ratio $\mathbf{R_{t\text{-}DS}}$ in table 5 is very close to one) while the G-DS procedure seriously underestimates the actual exceedance probabilities if tail dependence is present (distributions 3-8).

The G-DS underestimation is particularly severe when the Kendall's tau correlation of the original data is small (see distributions 3, 5, 7). That also should be expected, since under the Gaussian copula model, low Kendall's tau correlation is almost equivalent to independence. Consequently, the G-DS procedure shuffles the original variables as if they were independent thereby losing the dependence on the tails of the joint distribution. For distributions 5 and 7, characterized by tail dependence and low Kendall's tau correlation, the G-DS procedure, on average, underestimates the actual exceedance probabilities by a factor that ranges between 22 and 71 ($22.8<\mathbf{R_{G\text{-}DS}}<71.14$) while the t-DS procedure provides values of the $\mathbf{R_{t\text{-}DS}}$ ratio very close to 1. To better understand the consequences of such underestimation suppose that the variables (X,S) in the simulation study represent two types of insurance claims and that the distribution of (X,S) is distribution 7 (for a real data example where t copula is a suitable model for insurance claims see for example Resti et al. 2010). The estimated probability that on any day the two claims would follow above the 99.5% percentile of their margins (evaluated as the average of the query $Q_{2\alpha}$ over 1000 simulations) would be of 0.00129 for the original data, 0.00002 for the G-DS data and 0.00143 for the t-DS data. The differences in the responses have important practical consequences. The above results can be stated in the form of the query

Find how often in the long run both claims follow above the 99.5% percentile of their margins.

According to the original data based on the above results, the response to this query would be "once every 2.1 years" , the response using the t-DS data would be "once every 1.90 years" (quite close to the "true" value), while the response using the G-DS data would be "once every 115 years". Thus, with the DS shuffled data we would conclude that this event is 54.27 times less likely which would have significant consequences in decisions made using this data.

Also as expected the performance of the t-DS procedure deteriorates when we move to the asymmetric tail dependence case (λ_L very different than λ_U). We observe, for example, that for distributions 7-8 that show a moderate asymmetry in tail dependence, the t-DS procedure performs quite reasonably. However for more asymmetric distributions (as 9-12) the t-DS procedure performs poorly. For distributions 9-12 which exhibits lower tail dependence but negligible upper tail dependence, the t-DS procedure underestimates the probabilities that both variables take values below the α quantile of their margins up to a factor of 2.1 (query $Q_{1\alpha}$) and overestimates the upper tail dependence probabilities that both variables take values above the $1\text{-}\alpha$ quantile of their margins (query $Q_{2\alpha}$) up to a factor of 33. The symmetry implied by the t-copula when asymmetry is present leads to symmetry in the answers to queries $Q_{i\alpha}$ (i=1,2) under the t-DS procedure. The (lower and upper) tail dependence implied by t-DS procedure is "an average" of the lower and upper tail dependence in the data. When these two are very different the estimated tail dependence under the t-DS model is an incorrect representation of both (as it is the case for distributions 9-12).

In addition to the exceedance probabilities we also looked at higher order moments. In table 6 we report the average of the measures of multivariate skewness and kurtosis defined by Mardia et al. (1979, p.21) evaluated using the original data (and labeled as $\mathbf{A_{b1}}$ and $\mathbf{A_{b2}}$ respectively). The ratio between $\mathbf{A_{bi}}$ (i=1,2) and the corresponding value evaluated using the G-DS and t-DS shuffled data (ratios that we denote by $\mathbf{R_{bi}}$) are also reported in table 6, together with the

average biases ($\mathbf{AB_{b1}}$ and $\mathbf{AB_{b2}}$) and standard deviation of the bias for the two procedures ($\mathbf{SD_{Bb1}}$ and $\mathbf{SD_{Bb2}}$).

Distribution	1	2	3	4	5	6	7	8	9	10	11	12
Ab_1 original	0	0	0,01	0,01	0,01	0,01	0,02	0,01	0,05	0,05	0,23	0,28
R_{b1} G-DS	1,01	0,92	1,19	1,23	2,64	2,95	3,47	2,35	9,85	10,29	50,1	52,4
R_{b1} t-DS	0,99	0,91	1,01	0,95	1,57	1,29	2,02	1,26	6	7,59	30,04	44,26
AB_{b1} G-DS	0,04	-0,41	0,94	1,19	7,71	9,87	12,05	7,11	42,92	48,15	226,58	272,11
AB_{b1} t-DS	-0,03	-0,46	0,05	-0,32	4,52	3,36	8,54	2,52	39,8	46,31	223,5	271,14
SD_{Bb1} G-DS	3,45	4,07	4,27	4,78	10,11	14,79	14,65	8,61	25,29	17,55	42,99	36,73
SD_{Bb1} t-DS	3,44	3,78	5,03	6,21	11,09	17,54	15,47	11,33	25,84	17,73	43,45	37,02
Ab_2 original	8	8	8,4	8,5	10,9	10,6	10,9	8,9	10,5	8,5	9,6	8,5
R_{b2} G-DS	1,01	1	1,05	1,06	1,36	1,32	1,37	1,12	1,32	1,06	1,21	1,06
R_{b2} t-DS	1	1	1	1	1	1	1,01	0,96	0,97	0,99	0,97	1,01
AB_{b2} G-DS	0	0	0,4	0,5	2,9	2,6	3	0,9	2,5	0,5	1,7	0,5
AB_{b2} t-DS	0	0	0	0	0	0	0,1	-0,3	-0,4	-0,1	-0,3	0,1
SD_{Bb2} G-DS	0,1	0,1	0,1	0,1	0,1	0,3	0,1	0,1	0,1	0,1	0,1	0,1
SD_{Bb2} t-DS	0,1	0,1	0,1	0,2	0,1	0,4	0,1	0,3	0,1	0,2	0,1	0,2

*Table 6:**Skewness and Kurtosis for original, G-DS and t-DS shuffled data.***

Again we observe that G-DS and t-DS procedure performs equally well when no tail dependence is present (distributions 1 and 2); t-DS outperforms G-DS in the presence of tail dependence and null or moderate departure from symmetry in the tail dependence coefficients (distributions 3-8). However, both t-DS and G-DS perform poorly when the distribution underlying the original data (distributions 9-12) has very different lower and upper tail dependence (distributions 9-12).

6. **Conclusions**

The objective of this paper is to provide an alternative data shuffling method in the presence of tail dependence. The original data shuffling procedure was based on multivariate normal copulas and is not capable of maintaining tail dependence. The alternative data shuffling procedure (t-DS) presented in this paper that is based on t copulas has all the same characteristics as the original DS procedure but provides the additional advantage that it allows to model tail dependence as well, when the tail dependence has none to moderate asymmetry. The proposed new method might have important applications in disclosure limitation problems where dissemination of economic or actuarial data (for which tail dependence is a characteristic that often plays an important role in decision making) is the primary concern.

There are some disadvantages to using the t-DS procedure. First, tail dependence in the t-DS method is based on the single parameter ν and consequently all relationships are modeled using this single parameter. For some data, the degree of tail dependence may vary across the variables. In such cases, alternative methods for modeling the data will need to be devised. Second, the t-DS approach (or for that matter, all elliptical copulas based approaches) assumes that tail dependence is symmetric. In data sets where tail dependence is not symmetric, it may be necessary to consider alternative models as well (as we showed in section 5). These offer avenues for future research. Recent works on asymmetric copulas (Leibscher 2008) or grouped t copulas (Demarta and McNeil 2005) may provide a direction in this regard.

The general copula family also offers several avenues for future research in this area. Our discussion in this paper has been limited to continuous numerical data. For categorical data, existing masking methods that preserve margins (especially

those related to log-linear models), seem the natural alternative to the copula approach (see Fienberg and Slavkovic 2008 for an extensive discussion). For the general case of mixed data (binary, ordinal, and continuous) a copula approach based on the method of semiparametric inference for copula models via the extended rank likelihood (see Hoff 2007) is a promising tool for data masking and would be an interesting extension of the t-DS procedure to diverse data types.

Acknowledgements

This work was partly funded by the Spanish Ministry of Education and Science under the research project AYA2009-07981.

References

Arslan, O., 2004. Family of multivariate generalized t distributions. Journal of Multivariate Analysis, 89, 329-337.

Burridge, J., 2003. Information preserving statistical obfuscation. Statistics and Computing, 13, 321-327 .

Caillault, C. and Guégan D., 2005. Empirical estimation of tail dependence using copulas: application to Asian markets. Quantitative Finance, 5, 489 – 501.

Demarta, S. and McNeil, A. J., 2005. The t copula and related copulas. International Statistical Review, 73, 111-129 .

Fienberg, E. S. and Slavkovic A. B., 2008. A survey of statistical approaches to preserving confidentiality of contingency tables. In C. C. Aggarwal and P.S. Yu, eds., Privacy-Preserving Data Mining:models and Algorithms, 291-312.

Frees, E. W. and Valdez E. A.,1998. Understanding relationships using copulas. North American Actuarial Journal 2, 1-25.

Hoff, P. D., 2007. Extending the rank likelihood for semiparametric copula estimation. The Annals of Applied Statistics, 1, 265-283.

Joe, H., 1997. Multivariate models and dependence concepts, Chapman and Hall, London.

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.

Lee, S., Genton M. G., and Arellano-Valle, R. B., 2010. Perturbation of numerical confidential data via skew-t distributions. Management Science, 56, 318-333.

Leibscher, E., 2008. Construction of asymmetric multivariate copulas. Journal of Multivariate Analysis, 99, 2234-2250.

Muralidhar K., Parsa, R. and Sarathy R., 1999. A general additive data perturbation method for database security. Management Science, 45, 1399-1415.

Mardia, K.V., Kent, J. T., and Bibby, J. M., 1979. Multivariate Analysis. Academic Press London.

Muralidhar, K. and Sarathy R., 2003. A Theoretical Basis for Perturbation Methods. Statistics and Computing, 13, 329-335.

Muralidhar, K. and Sarathy R., 2006. Data shuffling - A new masking approach for numerical data. Management Science, 52, 658-670.

Nelsen, R. B. 2006: An Introduction to Copulas, 2nd edition. Springer, New York.

Patton. A., 2006. Modelling asymmetric exchange rate dependence. International Economic Review, 47, 527-556.

Resti Y., Ismail N. Jaaman S.H., 2010. Handling the dependence of claim severities with copula models. Journal of Mathematics and Statistics, 6, 136-142.

Rubin, D.B., 1993. Discussion on 'Statistical Disclosure Limitation'. Journal of Official Statistics, 9, 461-468.

Sarathy, R., Muralidhar K., and Parsa R., 2002. Perturbing non-normal confidential variables: The copula approach. Management Science, 48, 1613-1627.

Traub, J.F., Yemini Y., and Wozniakowski H., 1984. The statistical security of a statistical database. ACM Transactions on Database Systems, 9, 672-679.

Trottini, M., Muralidhar K., and Sarathy R., 2008. A preliminary investigation of the impact of Gaussian versus t-copula for data perturbation. In Privacy in Statistical Databases, Eds: Domingo-Ferrer and Saygin, Lecture Notes in Computer Science, 5262, Springer, Heidelberg, 127-138.