NISS

Database Security and Confidentiality: Examining Disclosure Risk vs. Data Utility through the R-U Confidentiality Map

George T. Duncan, Sallie A. Keller-McNulty, and S. Lynne Stokes

Technical Report Number 142
March, 2004

National Institute of Statistical Sciences
19 T. W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709-4006
www.niss.org

Database Security and Confidentiality: Examining Disclosure Risk vs. Data Utility through the R-U Confidentiality Map

George T. Duncan . Sallie A. Keller-McNulty . S. Lynne Stokes

Heinz School of Public Policy and Management, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213
Statistical Sciences Group, Los Alamos National Laboratory, Los Alamos, New Mexico 87545
Department of Statistics, Southern Methodist University, Dallas, Texas, 75275

gd17@andrew.cmu.edu . sallie@lanl.gov . slstokes@mail.smu.edu

---

Managers of database security must ensure that data access does not compromise the confidentiality afforded data providers, whether individuals or establishments. Recognizing that deidentification of data is generally inadequate to protect confidentiality against attack by a data snooper, managers of information organizations (IOs)—such as statistical agencies, data archives, and trade associations—can implement a variety of disclosure limitation (DL) techniques—such as topcoding, noise addition and data swapping—in developing data products. Desirably, the resulting restricted data have both high data utility U to data users and low disclosure risk R from data snoopers. IOs lack a framework for examining tradeoffs between R and U under a specific DL procedure. They also lack systematic ways of comparing the performance of distinct DL procedures. To provide this framework and facilitate comparisons, the *R-U confidentiality map* is introduced to trace the joint impact on R and U to changes in the parameters of a DL procedure. Implementation of an R-U confidentiality map is illustrated in the case of multivariate noise addition. Analysis is provided for two important multivariate estimation problems: a data user seeks to estimate linear combinations of means and to estimate regression coefficients. Implications for managers are explored.

(*Database Security; Multivariate Additive Noise; Confidentiality Protection; Disclosure Limitation; R-U Confidentiality Map; Topcoding*)

1. The Information Organization's Inferential Data Security Predicament

Information organizations (IOs) are organizations that capture, store, integrate and disseminate information. Examples are statistical agencies (U.S. Bureau of the Census), data archives (National Data Archive for Child Abuse and Neglect at Cornell), trade associations (National Association of Manufacturers), credit bureaus (Experian), marketing data firms (Nielsen//NetRatings), and health insurance information agencies (Health Insurance Industry Benchmarking Association**)**. IOs gather data from entities—individuals, households, and establishment—under implicit and explicit pledges to provide appropriate database security in giving access to legitimate users of the data (Schlörer 1981). They store these data and integrate them with other data. Importantly, they supply researchers and analysts with suitable data products. Overall, this can be described as the CSID (Capture Storage Integration Dissemination) data process (Duncan 2003). Managers of an IO need to assure inferential security in this CSID database system. These managers often are deterred from providing certain useful data products to their clients because they cannot ensure the security of a data product to an inference attack by a data snooper (Garfinkel, Gopal and Goes 2002). In that case the confidentiality of the data would be compromised, and the reputation of the IO as an honest broker damaged. Confidentiality protection may be required by law or regulation, as it is for the U.S. Census Bureau under Title 13 of the U.S. Code, or for health care organizations under the HIPAA Privacy Rule. Confidentiality may also be pledged—explicitly or implicitly—to a data provider, as it is for most surveys. A data product for which confidentiality could reasonably be compromised by data snooper attack has high disclosure risk.

Traditionally, the data products at most risk have been highly disaggregated tables with multiple attributes and fine resolution on each attribute scale. The cell entries would represent characteristics of just a few entities. For example, suppose Forrester Research contemplated

releasing to clients a two-way table of total corporate R & D expenditures, cross-classified so rows would give geography to the county level and columns would give product to a fairly detailed level. The cell in this table corresponding to ("King County, WA", "ERP Software") would likely include several firms but be dominated by Microsoft's expenditures, and hence provide sensitive information to competitors about their strategic commitment to ERP software. We would call this an *attribute disclosure*, as information has been disclosed about an attribute of an entity (Duncan and Lambert 1989).

Increasingly, analysts want the even richer data products represented by microdata, which involve compilations of individual records. For example, the BlueCross BlueShield Association might like access to microdata on healthcare utilization of individual HMO subscribers together with their demographic attributes; the HMO through its corporate structure, co-payment requirements, etc.; and subscriber's employer through its nature of business, size, and health insurance alternatives made available to employees. Because of the obvious richness of such a record, deidentification—stripping apparent identifiers such as name, SSN, email address, etc—would not prevent reidentification through linkage with a variety of external databases that included identifiers, notably marketing databases. With such reidentification, the data snooper would have successfully attacked the database and there would have been an inferential disclosure that compromised confidentiality. Specifically, we call this an *identity disclosure* (Duncan and Lambert 1989), since a specific identity is linked to a record in the purportedly protected database.

With the popularity of the web, data users want IOs to provide flexible, online access to data resources. To some extent, IOs have been responsive to this demand, as the U. S. Census Bureau now provides some access to user-specified tables through American FactFinder. But concerns about confidentiality disclosure risk continue to inhibit the growth of this access mode.

IOs face what computer scientists call the *Inference Problem*, which is the deduction of confidential data from nonsensitive data objects (Keller-McNulty and Unger 1993). The IO's

predicament is that, with an inappropriate choice of DL procedure or parameter value of a DL procedure, the utility of their data product may suffer unduly or disclosure risk may yet remain too high. Present practice by IOs in assessing tradeoffs between disclosure risk and data utility is largely heuristic, and so would benefit from an appropriate theoretical framework. Indeed, Recommendation 6.2 of the National Academy of Sciences Panel on Confidentiality and Data Access (Duncan, Jabine and de Wolf 1993) urges the development of foundations for the analysis of tradeoffs between disclosure risk and data utility

Here is an example of the concern by an information organization for disclosure limitation:

> The Health and Retirement Study, conducted by the University of Michigan under funding from the National Institute on Aging, promises, "All answers are treated as strictly confidential." Record linkage of the survey results with earnings and benefits data from the Social Security Administration (SSA) adds much to the data's utility but increases disclosure risk. For a discussion of this concept, see Duncan and Lambert (1989), Lambert (1993), and Elliot and Dale (1999). A variety of methods, including removing geographic information, rounding, and top-coding, were used to lower risk of disclosure (http://micda.psc.isr.umich.edu/enclave/DisclosureReview.pdf)

In all such examples, the key question is whether the disclosure limitation methods used are adequate, but not excessive. Could less severe distortion or obscuring of the data still keep low the risk from data snoopers, while allowing better data utility? What explicitly is the tradeoff between disclosure risk and data utility? Would a different DL method lower disclosure risk while maintaining data utility?

In this article, we explain in Section 2 how the R-U confidentiality map helps managers determine the form in which access can be provided to their data. We show in Section 3 that the R-U map can be computed in useful microdata cases, illustrated by the DL technique of multivariate noise addition. We assume throughout that the snooper strategy is identification of

the target record index through matching to an external file. In Section 4, we derive expressions for R and U for three multivariate estimation problems: (1) linear combination of means of masked variables, (2) regression coefficients for a simple linear regression of an unmasked variable on a masked one; and (3) regression coefficients for simple linear regression of a masked variable on an unmasked one. In Section 5, we show how these analytic expressions can be used for making managerial decisions about how to protect the IO's data. In Section 6 we draw conclusions, specify some of the advantages of the R-U confidentiality map for managerial practice, and identify promising extensions.

2. A Managerial Analysis: The R-U Confidentiality Map

An R-U confidentiality map provides a useful analytical framework for managers of IOs to assess tradeoffs between the benefits of providing data products and the risks involved in doing so. An IO brokers information from data providers to data users. The data users are legitimate clients of the IO; they follow the rules and pose no harm to the promises the IO has made to data providers. A data snooper, on the other hand, does seek information about individual data providers that could compromise those promises. The manager's job is to develop and maintain policies to ensure that valuable data flows to the clients while stymieing any efforts by the data snooper to compromise the confidentiality of the database. As with most security problems there are tradeoffs in trying to achieve these two goals. We can protect something valuable by locking it away, but then it has no use. We can make it freely available for use, but then it is vulnerable. A decision-theoretic framework for modeling a data snooper's behavior is given by Duncan and Lambert (1986). A comprehensive framework for the IO manager's problem is given by Trottini (2001, 2003). With the R-U confidentiality map, the IO manager has an analytic tool for systematically examining tradeoffs between value to a client, *data utility*, and vulnerability to a data snooper, *disclosure risk*. A beginning on this task in the form of a very basic *R-U*

confidentiality map is presented in Duncan and Fienberg (1999) in the context of tabular data. A further development, again in the context of tabular data, is given in Duncan *et al* (2001). An application in the healthcare context is provided by Boyens, Krishnan and Padman (2004).

A measure of statistical disclosure risk, R, is a numerical assessment of the risk of unintended disclosures to a data snooper from dissemination of the data product. A measure of data utility, U, is a numerical assessment of the usefulness of the released data to legitimate users. When this utility U is based on the discrepancy between the masked data and the original data, it is called a *distortion measure* (Gomatam and Karr 2003).

An IO can lower the disclosure risk of a data product by applying a disclosure limitation (DL) procedure to mask the data (Agrawal and Aggarwal 2001; Duncan 2002; Keller-McNulty and Unger 1993; Muralidhar, Batrah and Kirs 1995; Muralidhar, Parsa and Sarathy 1999; Muralidhar and Sarathy 1999; Tendick and Matloff 1994). Disclosure limitation involves stochastic or deterministic transformations of the original data or the generation of synthetic data using a model constructed on the basis of the original data (Raghunathan, Reiter and Rubin 2003). DL transformations are designed to hamper a data snooper in reidentifying a record or learning sensitive information about a particular individual or enterprise. For a general exploration of confidentiality and data access issues, see Duncan, Jabine and de Wolf (1993). The literature in disclosure limitation includes Adam and Wortman (1989), Duncan and Pearson (1991), Fienberg (1994), Gopal, Goes, Garfinkel (2002), Jabine (1993), Kooiman, Nobel and Willenborg (1999), Marsh *et al* (1991), Mackie and Bradburn (2000), and Willenborg and de Waal (1996). Because this masking, while intended to lower disclosure risk, will typically also lower data utility, it is crucial that IOs assess the tradeoff.

The R-U confidentiality map traces the joint impact on R and U of changes in parameter values of the DL procedure, thereby enabling comparison of DL procedures and tradeoffs between disclosure risk and data utility. In the next section, we illustrate how an R-U

confidentiality map can be constructed for an important class of DL procedures—multivariate additive noise.

3. Constructing an R-U Confidentiality Map: Multivariate Additive Noise

In this section, we outline how to construct and use R-U confidentiality maps for a particular disclosure limitation method—multivariate additive noise. Additive noise has been proposed and examined as a disclosure limitation device by several authors. These include, in the univariate case, Brand (2002), Spruill (1983), Paass (1988), and Duncan and Mukherjee (2000), and in the multivariate case, Sullivan and Fuller (1989). We follow an implementation discussed by Kim (1986), and explored by Kim and Winkler (1995). It has been implemented by the U.S. Census Bureau (Moore 1996). As noted in Zayatz, Moore and Evans (1996), "To employ it effectively, the user must exhibit some expertise in the setting of various parameters." We demonstrate how the R-U confidentiality map provides an effective decision aid for this task.

Our structure and notation is as follows: The original data is $\underset{n\times p}{\mathbf{X}} = [X_{ij}] = [\mathbf{X_1},\ldots,\mathbf{X_n}]'$ where $\mathbf{X_i} \overset{iid}{\sim} (\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The first p_1 variables in the database are deemed sensitive and the remaining p_2 variables are deemed non-sensitive, with $p = p_1 + p_2$. The covariance matrix $\boldsymbol{\Sigma}$ is partitioned according to the sensitivity of the fields, as

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma_{11}} & \boldsymbol{\Sigma_{12}} \\ \boldsymbol{\Sigma_{21}} & \boldsymbol{\Sigma_{22}} \end{bmatrix}.$$

Sensitive variables are masked by additive noise, independent of $\mathbf{X}$, that has the same correlation structure as the original data. That is, writing the data after masking as **Y**, we have

$$\mathbf{Y} = \mathbf{X} + \boldsymbol{\varepsilon} \qquad (3.1)$$

where $\boldsymbol{\varepsilon}' = (\boldsymbol{\varepsilon}_1', \boldsymbol{\varepsilon}_2')$ with $\underset{p_1 \times 1}{\boldsymbol{\varepsilon}_1} \sim (\mathbf{0}, \lambda^2 \boldsymbol{\Sigma}_{11})$ and $\boldsymbol{\varepsilon}_2$ a $p_2 \times 1$ vector of 0's. The DL parameter λ^2 adjusts the extent of masking. The advantage of this approach over addition of independent noise to each sensitive variable is that the correlation structure among the p_1 masked sensitive variables is the same as that of the original sensitive variables. The correlation structure among the p_2 unmasked non-sensitive variables is, of course, also unchanged. The correlation between a masked sensitive variable and an unmasked non-sensitive variable will be attenuated by a factor of $\frac{1}{\sqrt{1+\lambda^2}}$.

Data users are interested in statistical inference about parameters of the distribution of **X**. To be specific in our discussion, we consider a prototypical data user interested in estimating some parameter θ of the distribution of **X** who uses the same estimator of θ as if he had access to **X**, but will actually compute from the released data **Y**. We denote the estimator that would be computed from the unmasked data as $\hat{\theta}_x$ and the estimator the data user computes from the masked data as $\hat{\theta}_y$. For this illustration, we take the data utility U to be the reciprocal of the data user's mean squared error; that is,

$$U = [E(\hat{\theta}_y - \theta)^2]^{-1} . \qquad (3.2)$$

In contrast to the data user's statistical interest in θ, the data snooper has a confidentiality-compromising interest in the value τ of one of the sensitive variables for an individual entity. The snooper predicts that value from the released data. This prediction is denoted by $\hat{\tau}$. We assume that the data snooper uses record linkage to match the released database to an outside source of data using some unmasked key variables, and then takes $\hat{\tau}$ to be the value of the sensitive variable in the matching record for the targeted individual. If the sensitive variable has been

masked, $\hat{\tau}$ may differ from τ, even if the match is accurate. We take the disclosure risk *R* to be the reciprocal of the data snooper's target mean squared error, that is

$$R = [E(\hat{\tau} - \tau)^2]^{-1} . \qquad (3.3)$$

In practice, the form of R and U should be tailored to the particular situation at hand. For example, if management's concern was identity disclosure rather than attribute disclosure, risk could be defined as the expected proportion of records in the (masked) file correctly identified by the data snooper. If the masking process was deterministic (e.g., topcoding, in which all values above some threshold are replaced by a single value), then $\hat{\tau}$ will not be random, and *R* would not be defined as the expectation over all possible masked datasets as in (3.3), but instead could be defined as $R = [(\hat{\tau} - \tau)^2]^{-1}$.

For any specific data user, parameter θ, and snooper strategy, the disclosure risk *R* and data utility *U* are functions of the parameter(s) of the disclosure limitation method. In the case of noise addition, both are functions of the noise multiple λ^2. The R-U confidentiality map is then constructed by plotting the pairs (*U*, *R*) for the range of values of λ^2 under consideration.

In Section 4, we derive these functions for several types of parameters θ: (1) linear combination of means of masked variables $\theta = \mathbf{c}'\boldsymbol{\mu}_1$; (2) regression coefficients for a simple linear regression of an unmasked variable on a masked one; and (3) regression coefficients for a simple linear regression of a masked variable on an unmasked one. We assume throughout that the snooper strategy is identification of the target record index through matching to an external file. In Section 5, we show how these analytic expressions can be used for making managerial decisions about how to protect the IO's data.

4. *R* and *U* for Multivariate Estimation

As specified in Equations 3.2 and 3.3, respectively, *U* and *R* can be written as functions of λ^2 under model (3.1). They will take different forms depending upon what parameters are of interest to the data user and what strategies the snooper employs for predicting the target attribute value. In this section we examine special cases that are multivariate and of practical importance.

In Case 1 we take the data user's purpose to be estimation of a linear combination of the means of the sensitive variables, so $\theta = \mathbf{c}'\boldsymbol{\mu}_1$. These variables have been masked in the released file and the user computes $\hat{\theta}_y = \mathbf{c}'\overline{\mathbf{Y}}_\mathbf{1}$ as an estimator of θ, where $\overline{\mathbf{Y}}_\mathbf{1}$ is the sample mean vector of the masked variables. If the database is a simple random sample of size *n* from the population of interest, then $E(\hat{\theta}_y) = \theta = \mathbf{c}'\boldsymbol{\mu}_\mathbf{1}$ and $Var(\hat{\theta}_y) = \frac{(1+\lambda^2)(\mathbf{c}'\boldsymbol{\Sigma}_{\mathbf{11}}\mathbf{c})}{n}$. Then from (3.2) we have the data utility as

$$U_{c'\mu} = \frac{n}{(1+\lambda^2)(\mathbf{c}'\boldsymbol{\Sigma}_{\mathbf{11}}\mathbf{c})}. \qquad (4.1)$$

Next we take the data user's goal to be estimation of the slope coefficient β of the simple linear regression model

$$X_{r'} = \alpha + \beta X_r + e, \qquad (4.2)$$

where $e \sim N(0, \sigma_e^2)$. In Case 1 where both the independent and dependent variables in model (4.2) are masked using model (3.1), we now demonstrate that estimation of the regression coefficient is unaffected. We then examine Cases 2 and 3 where masking is only on the independent variable or dependent variable.

Case 1. We show that the mean and variance of the least squares simple linear regression slope coefficient is unchanged when both original independent and dependent variables, X_r and $X_{r'}$ are masked with noise having the same covariance structure as X_r and $X_{r'}$. Beyond model (4.2),

we take $X_r \sim \text{Normal}(\mu_r, \sigma_r^2)$, so that the vector $(X_r, X_{r'})$ has a bivariate normal distribution. Specifically, $(X_r, X_{r'})' \sim \text{BivNorm}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where

$\boldsymbol{\mu} = \begin{pmatrix} \mu_r \\ \mu_{r'} = \alpha + \beta\mu_r \end{pmatrix}$ and $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_r^2 & \sigma_{r,r'} = \beta\sigma_r^2 \\ \beta\sigma_r^2 & \sigma_{r'}^2 = \beta^2\sigma_r^2 + \sigma_e^2 \end{pmatrix}$. Now we mask both X_r and $X_{r'}$ as $Y_r = X_r + \varepsilon_r$ and $Y_{r'} = X_{r'} + \varepsilon_{r'}$, where

$(\varepsilon_r, \varepsilon_{r'})' \sim \text{BivNorm}((0,0)', \lambda^2\boldsymbol{\Sigma})$ (4.3)

and independent of $(X_r, X_{r'})$. Then $(Y_r, Y_{r'})' \sim \text{BivNorm}(\boldsymbol{\mu}, (1+\lambda^2)\boldsymbol{\Sigma})$. Note that the correlation between Y_r and $Y_{r'}$ is the same as that between X_r and $X_{r'}$, and that $1-\rho^2 = \sigma_e^2 / (\beta^2\sigma_r^2 + \sigma_e^2)$. Now suppose we estimate the regression model (4.2) using the masked data. Then the least squares estimator of the regression parameter vector $\boldsymbol{\gamma}' = (\alpha, \beta)$ is $\hat{\boldsymbol{\gamma}} = \mathbf{(Z'Z)^{-1}Z'Y_{r'}}$, where $\mathbf{Z} = \mathbf{[1, Y_r]}$ is the usual design matrix for a simple linear regression, where $\mathbf{Y_r}$ is the vector of masked independent variables and $\mathbf{Y_{r'}}$ is the vector of masked dependent variables. This estimator is still unbiased, since $\mathbf{E}(\hat{\boldsymbol{\gamma}}) = \mathbf{E}\mathbf{E}(\hat{\boldsymbol{\gamma}} \mid \mathbf{Y_r}) = \mathbf{E}(\boldsymbol{\gamma}) = \boldsymbol{\gamma}$. To calculate the variance, observe that

$\mathbf{Var}(\hat{\boldsymbol{\gamma}}) = \mathbf{E}[\mathbf{Var}(\hat{\boldsymbol{\gamma}} \mid \mathbf{Y_r})] + \mathbf{Var}[\mathbf{E}(\hat{\boldsymbol{\gamma}} \mid \mathbf{Y_r})] = \mathbf{E}[\mathbf{(Z'Z)^{-1}Z'Var(Y_{r'} \mid Y_r)Z(Z'Z)^{-1}}]$

$+ \mathbf{Var}(\boldsymbol{\gamma})$. The second term is 0 since the conditional expectation of $\hat{\boldsymbol{\gamma}}$ does not depend upon $\mathbf{Y_r}$. To calculate the first term, note that $\mathbf{Var(Y_{r'} \mid Y_r)} = (1-\rho^2)(1+\lambda^2)\sigma_{r'}^2\mathbf{I} = (1+\lambda^2)\sigma_e^2\mathbf{I}$, yielding $\mathbf{Var}(\hat{\boldsymbol{\gamma}}) = (1+\lambda^2)\sigma_e^2\mathbf{E(Z'Z)^{-1}}$. We are interested in the variance of the slope estimator, so we write the $(2,2)^{\text{th}}$ term of $\mathbf{E(Z'Z)^{-1}}$ as $E\left\{\left[\sum (Y_{ri} - \overline{Y}_r)^2\right]^{-1}\right\}$

$= \left[(n-3)(1+\lambda^2)\sigma_r^2\right]^{-1}$ since the random variable in curly brackets has (a multiple of) an inverse chi-square distribution with n-1 degrees of freedom. Thus

$$Var(\hat{\beta}) = \frac{(1+\lambda^2)\sigma_e^2}{(n-3)(1+\lambda^2)\sigma_r^2} = \frac{\sigma_e^2}{(n-3)\sigma_r^2}.$$

This is the same expression for the variance had the estimator used the unmasked data. Inference problems occur, however, when only one of the two variables involved in the regression model (4.2) is masked. We consider these two situations next. In each case, we again suppose the data user computes the usual least squares estimator from the released data in the usual way.

Case 2. The independent variable X_r is a masked variable and the dependent variable $X_{r'}$ is not (Case 3 is vice versa.) Case 2 is well-studied in the measurement error literature. When the independent variable is noisy, the estimator of the slope in a simple linear regression is biased toward 0, or attenuated. Specifically, assuming the regression model (4.2), the model for the masking variable (4.3), and normality of the independent variable X_r, the estimator $\hat{\beta}_2 = \sum_i (X_{r'i} - \overline{X}_r)(Y_{ri} - \overline{Y}_r) / \sum_i (Y_{ri} - \overline{Y}_r)^2$ has mean

$$E(\hat{\beta}_2) = \beta / (1+\lambda^2) \qquad (4.4)$$

and variance

$$Var(\hat{\beta}_2) = \frac{\beta^2\sigma_r^2(1 - \frac{1}{1+\lambda^2}) + \sigma_e^2}{(n-3)\sigma_r^2(1+\lambda^2)}. \qquad (4.5)$$

(See Fuller (1987), Equations (1.1.6) and (1.1.11), respectively.) From (3.2), (4.4), and (4.5), we compute utility to be

$$U_{\hat{\beta}_2} = \left\{ \frac{\beta^2\lambda^2}{(1+\lambda^2)^2}\left(\frac{1}{n-3} + \lambda^2\right) + \frac{\sigma_e^2}{(n-3)(1+\lambda^2)\sigma_r^2} \right\}^{-1}$$

$$= \left\{ \frac{\beta^2}{(1+\lambda^2)} \left[\frac{\lambda^2}{(1+\lambda^2)} \left(\frac{1}{n-3} + \lambda^2 \right) + \frac{1}{n-3} \frac{1-R^2}{R^2} \right] \right\}^{-1} \qquad (4.6)$$

where

$$R^2 = 1 - \frac{\sigma_e^2}{\sigma_{r'}^2} = 1 - \frac{\sigma_e^2}{\beta^2 \sigma_r^2 + \sigma_e^2}. \qquad (4.7)$$

Note that the utility is an increasing function of sample size (n) and a decreasing function of λ^2, as we would expect. The shape of the utility curve depends on β only indirectly through R^2, a measure of the strength of the linear relationship between X_r and $X_{r'}$.

Case 3. The masked variable is the dependent rather than the independent variable. Again the data user estimates the slope coefficient β in model (4.2), but now uses the estimator

$$\hat{\beta}_3 = \Sigma_i (Y_{r'i} - \overline{Y}_{r'})(X_{ri} - \overline{X}_r) \big/ \Sigma_i (X_{ri} - \overline{X}_r)^2 .$$

When the error is in the dependent variable, there is no attenuation of the slope estimator, but its variance is increased. Under the model described above, the variance of this estimator is

$$Var(\hat{\beta}_3) = \frac{\lambda^2 (\beta^2 \sigma_r^2 + \sigma_e^2) + \sigma_e^2}{(n-3)\sigma_r^2} .$$

(See Fuller (1987), Eq. (1.1.11).) This yields a utility of

$$U_{\hat{\beta}_3} = \left\{ \beta^2 \frac{(\lambda^2 + 1 - R^2)}{(n-3)R^2} \right\}^{-1}, \qquad (4.8)$$

where, R^2 is defined in (4.7). Note that this utility, too, is increasing in n, decreasing in λ^2, and has shape determined by β only through R^2.

Now we consider how to calculate disclosure risk R. In order to do so, we must model the snooper's strategy for gaining access to the database information. In this paper, we assume that the snooper has gained index knowledge. By this, we mean that the snooper can actually identify the record within the database that pertains to his target. The most likely method that the snooper

could use to obtain index knowledge is by matching the released data to identified external data on the basis of key variables, usually demographic ones, which are not masked. Such identified databases are now readily available. For example, the Republican National Committee has Voter Vault and the Democratic National Committee has DataMart, with names, addresses, voting histories, income ranges, ages, ethnicities, marital statuses and more on some 168 million registered voters in the U.S. With such record linkage capability, the snooper can take the value of his target variable to be that associated with the identified record. We denote the target value by X. If this variable has been masked with additive noise, then the disclosure risk from this snooper attack can be calculated from (3.3) as

$$R = \frac{1}{E(X - Y)^2} = \frac{1}{\lambda^2 \sigma^2}, \qquad (4.8)$$

where σ^2 = Var(X). If the snooper's linking capabilities are less than perfect, then the risk will be reduced. So we can consider the expression above to be a conservative assessment of disclosure risk.

5. Manager's use of the R-U Confidentiality Map

How can a manager use the R-U confidentiality map for making decisions about how to release microdata products? Specifically, suppose that a manager wants to release demographic microdata from a survey of teachers. The survey data include salary and household income. Both of these income variables are to be masked by multivariate noise addition as described in Section 3. The manager knows that users will be interested in estimating the mean of each variable, as well as the difference in the two. Because the database also contains numerous unmasked demographic variables, such as zip code and race, the manager reasonably anticipates that a data snooper could learn the identity (i.e., have index knowledge) thus the value of the protected variables for some respondents. Suppose that the manager would like to select the amount of noise to be added (λ^2) so that the snooper's expected error for the salary variable is at least \$5K.

This translates into a maximum tolerable risk (from (3.3)) of $R_{\max} = (\$5000)^{-1}$. Suppose that the estimated variance-covariance matrix $\mathbf{\Sigma_{11}}$ for the pair of variables X_1 = household income and X_2 = salary (in units of a thousand dollars) is

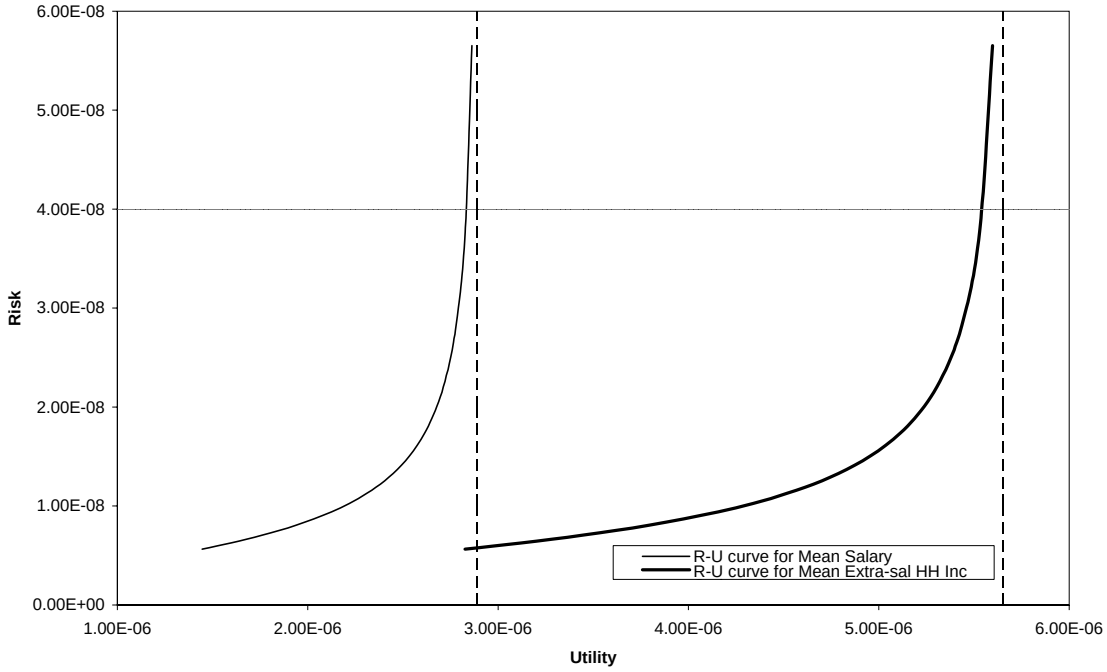
$$\begin{pmatrix} 17.5^2 & 8.3^2 \\ 8.3^2 & 13.3^2 \end{pmatrix},$$

and the survey data contains n = 1000 records. Then the data utility for the parameters of interest to the data users are, from (4.1), $U_{\mu_2} = \dfrac{1000}{(1+\lambda^2)(13.3)^2}$ and $U_{\mu_1-\mu_2} = \dfrac{1000}{(1+\lambda^2)(18.6)^2}$. The risk of disclosure for the salary attribute to a snooper with index knowledge is, from (4.8),

$$R = \frac{1}{\lambda^2 (13.3)^2}.$$

Figure 1 shows an R-U confidentiality map that would inform the manager about the appropriate magnitude of masking. The left-hand curve shows the trade-off between risk and utility for μ_2, the mean salary. The right-hand curve shows the trade-off between risk and utility for $\mu_1 - \mu_2$,, the mean additional (extra-salary) household income. The horizontal dashed line is at the maximum tolerable risk R_{max} as set by the manager. The vertical dashed lines show the utility for μ_2 and $\mu_1 - \mu_2$ if the data were not masked at all. The manager's requirement for adequate masking can be achieved for λ^2 approximately 0.15. Figure 1 shows that the loss in utility for this level of disclosure limitation has very little effect on the utility of the data for μ_2, but slightly more for $\mu_1 - \mu_2$. The value of λ^2 could be increased still further before a rapid decrease in utility for μ_2 occurs, but the same is not true for $\mu_1 - \mu_2$.

Figure 1. R-U Confidentiality Map for Salary and Extra-Salary Household Income



For our second example, the manager at an organization which provides secondary medical data is preparing to release microdata containing variables such as the extent of alcohol consumption and dollars spent on psychiatric drugs. Variables viewed as sensitive are to be masked using multivariate noise, but others deemed not to be sensitive will not be masked. The manager knows that users of their microdata commonly examine relationships among the variables, so would like to investigate the utility for their data users who plan estimation of slope coefficients in simple regressions. Multivariate noise addition was chosen as the masking method in this case because it has the property that neither bias nor variance of estimated regression coefficients are affected when both independent and dependent variables in a simple regression are masked. However, if only the independent variable is masked, then user goals may be seriously affected by bias. Typically, the masking procedures details, such as the additive noise variance multiplier, λ^2, are not released to avoid providing any more information than necessary. In this case, however, the manager is considering releasing its value so that sophisticated data

users can make attenuation adjustments. The decision will be based upon size of the resulting increase in U when λ^2 is large enough to protect the identity of those possessing extreme values.

The utility for the estimator of the slope coefficient β in model (4.2) is shown in (4.6). If the user knows the value of λ^2, however, the bias in $\hat{\beta}_2$ can be eliminated by using the adjusted estimator $\hat{\beta}^* = (1+\lambda^2)\hat{\beta}_2$. The utility for this estimator is

$$U_{\hat{\beta}^*} = \left[Var\,(\hat{\beta}^*)\right]^{-1} = \left[(1+\lambda^2)^2\, Var\,(\hat{\beta}_2)\right]^{-1}$$

$$= \left\{ \frac{\beta^2 \lambda^2}{(n-3)} + \frac{(1+\lambda^2)\sigma_e^2}{(n-3)\sigma_r^2} \right\}^{-1} \qquad (5.1)$$

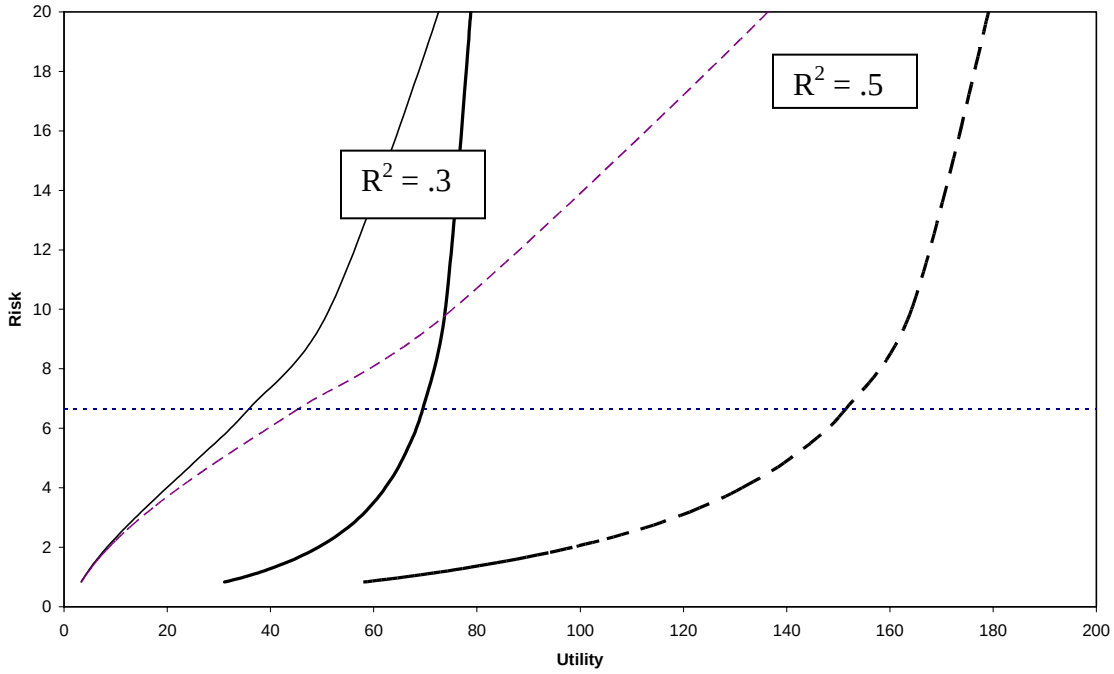
$$= \left\{ \beta^2 \frac{(\lambda^2 + 1 - R^2)}{(n-3)R^2} \right\}^{-1}$$

from (4.5). Note that this expression is identical to that for the utility for the adjusted estimator when only the dependent variable is masked, shown in (4.8).

To examine how much advantage the data user gains from knowing λ^2, the manager prepares R-U confidentiality maps comparing $U_{\hat{\beta}_2}$ (from (4.6)) and $U_{\hat{\beta}^*}$ (from (5.1)). Figure 2 shows such a map for a sample of n = 200, two values of R^2: 0.3 (solid lines) and 0.5 (dashed lines), and for risk calculated from (4.8). Without loss of generality, the values of β (in (4.6) and (5.1)) and σ (in (4.8)) were set to 1, since the comparison is unaffected by their values. The lighter (left-most) line in each pair shows $U_{\hat{\beta}_2}$, the analysts utility for $\hat{\beta}_2$; the darker (right-most) line shows $U_{\hat{\beta}^*}$ the utility for the slope estimator corrected for attenuation. It is clear from the figure that the gain in utility from knowledge of λ^2 is large for even moderately strong relationships. For example, with R^2 = .3, data utility would be approximately doubled if λ^2 =.15, shown as the dotted line on the figure. As the strength of the relationship increases, the loss in

data utility becomes even greater. Because of this large gain, the manager may decide it will be worthwhile to change policies to release of λ^2.

Figure 2. R-U map to Investigate Value of Data User's Knowledge of λ^2



6. Conclusions

This article develops the R-U confidentiality map as a tool that can help managers of information organizations make better decisions about how to use disclosure limitation (DL) methods. They can then better fulfill their dual mandate of providing useful data while maintaining an adequate level of database confidentiality. We showed how to construct the tool for the important DL method of multivariate noise addition and analyst goals of estimating linear combinations of means and regression coefficients.

An advantage to the IO of the process of developing an R-U confidentiality map for their own data and DL method is that it requires explicit formulations of R and U that are relevant to the needs of their communities of respondents (in thinking about R) and data users (in thinking about U), and their own institutional needs in thinking about both R and U. The IO may be

encouraged to monitor more closely, for example, what analyses are most frequently implemented on their data, or what estimated parameters are considered to be most important to their users, in order to more realistically model data utility. Information of this sort may become more easily available by monitoring queries to data websites with built-in analysis tools, as they are now becoming available. Information for modeling of risk can come from gathering data about perceived risks from disclosure that are of concern to their respondents. A data snooper attack could be simulated, say using administrative records (see Paass 1988).

There are several further avenues for development of the R-U confidentiality map that would be useful. First, we find value in developing analytical R-U maps for some more complex DL methods, such as data swapping (Dalenius and Reiss 1982) and the generation of synthetic or virtual data (Rubin 1993, Abowd and Woodcock 2001). The R-U confidentiality map itself could be generalized to address more complex decisions about a DL choice. For example, it might be useful to combine R-U maps for those cases where the utility of a variety of different parameter estimates must be considered, for example by plotting a weighted average or maximum of a small set of R and U values. Another example of an adaptation of the concept of an R-U confidentiality map would be one that allows exploration of the risk and utility tradeoff for a DL procedure indexed by two or more parameters, which is suggested by disclosure limitation through microaggegation and binning (Domingo-Ferrer and Torra 2001).

Overall, the R-U confidentiality map provides an appropriate conceptual framework for the information management task of protecting database security against snooper attack. It identifies the key elements of disclosure risk R and data utility U, provides an analysis of tradeoffs between R and U, and allows comparisons of different disclosure limitation (DL) methods.

Acknowledgments

This work was partially supported by grants from the National Science Foundation under Grant EIA-9876619 to the National Institute of Statistical Sciences, the National Center for Education Statistics under Agreement EDOERI-00-000236 to Los Alamos National Laboratory, and the National Institute on Aging under Grant 1R03AG19020-01 to Los Alamos National Laboratory. Initial work began under contract to George Duncan from the U.S. Census Bureau under Contract OBLIG-1999-17087-0-0 to Carnegie Mellon University. Later work was done at Los Alamos National Laboratory where George Duncan was on leave as a Visiting Faculty Member from Carnegie Mellon University and Lynne Stokes was a Visiting Faculty Member from Southern Methodist University. For helpful discussions on this topic, the authors wish to thank Alan Karr, Stephen Roehrig, and Karthik Kannon.

References

Abowd, J. M. and Woodcock, S. D. (2001) Disclosure limitation in longitudinal linked data. *Confidentiality, Disclosure and Data Access: Theory and Practical Application for Statistical Agencies* (Doyle, Lane, Theeuwes, and Zayatz, eds.) North-Holland 215-278.

Adam, N. R. and Wortman, J. C. (1989) Security-control methods for statistical databases: a comparative study. *ACM Computing Surveys* **21** 515-556.

Agrawal, D. and Aggarwal, C. C. (2001), "On the Design and Quantification of Privacy Preserving Data Mining Algorithms," *Proceedings of the Twentieth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems,* Santa Barbara, California, USA, May 21-23 2001.

Boyens, C., Krishnan, R. and Padman, R. (2004). On privacy-preserving access to distributed heterogeneous healthcare information. Proceedings of the 37th Hawaii International Conference on Systems Sciences.

Brand, R. (2002), "Microdata Protection through Noise Addition," in (J. Domingo-Ferrer, ed.) *Inference Control in Statistical Databases*, Springer: New York, 97-116.

Cramér, H. (1946) *Mathematical Methods of Statistics*, Princeton University Press.

Dalenius, T, and Reiss, S.P. (1982), "Data-swapping: A Technique for Disclosure Control," *Journal of Statistical Planning and Inference*, **6**, 73-85.

Domingo-Ferrer, J. and Terra, V. (2001) Disclosure control methods and information loss for microdata. *Confidentiality, Disclosure and Data Access: Theory and Practical Application for Statistical Agencies* (Doyle, Lane, Theeuwes, and Zayatz, eds.) North-Holland 91-110.

Duncan, G. T. (2002) Confidentiality and statistical disclosure limitation. *International Encyclopedia of the Social and Behavioral Sciences.* N. J. Smelser and Paul B. Baltes (editors) Pergamon, Oxford. 2521-2525.

Duncan, G. T. (2003**)** Exploring the Tension between Privacy and the Social Benefits of Government Databases, Security, Technology, and Privacy: Shaping a 21st Century Information Policy. The Century Foundation.

Duncan, G. T. and Fienberg, S. E. (1999) Obtaining information while preserving privacy: a Markov perturbation method for tabular data. Eurostat. *Statistical Data Protection '98* Lisbon 351-362.

Duncan, G. T., Fienberg, S. E., Krishnan, R., Padman, R. and Roehrig, S. F. (2001) Disclosure limitation methods and information loss for tabular data. In *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies* (Doyle, Lane, Theeuwes, and Zayatz, eds.) North-Holland 135-166.

Duncan, G. T., Jabine, T. B., and de Wolf, V. A. (1993) *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics.* Washington, D.C.: National Academy Press.

Duncan, G. T. and Lambert, D. (1989) The risk of disclosure for microdata. *Journal of Business and Economic Statistics* **7** 207-217.

Duncan, G. T. and Mukherjee, S. (2000). Optimal disclosure limitation strategy in statistical databases: deterring tracker attacks through additive noise. *Journal of the American Statistical Association* **95** 720-729.

Duncan, G. T. and Pearson, R. W. (1991) Enhancing access to microdata while protecting confidentiality. *Statistical Science* **6** 219-239.

Elliot, M. and Dale, A. (1999) Scenarios of attack: the data intruder's perspective on statistical disclosure risk. *Special issue on statistical disclosure control. Netherlands Official Statistics* **14** 6-10.

Fienberg, S. E. (1994) Conflicts between the needs for access to statistical information and demands for confidentiality. *Journal of Official Statistics* **10** 115-132.

Fisher, R.A. and Tippett, L.H.C. (1928) Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proceedings of the Cambridge Philosophical Society* **24** 180-190.

Garfinkel, R., Gopal, R., and Goes, P. (2002) Privacy protection of binary confidential data against deterministic, stochastic, and insider threat. *Management Science* **48** 749-764.

S. Gomatam and A. F. Karr (2003). Distortion measures for categorical data swapping. Technical Report, National Institute of Statistical Sciences. Research Triangle Park, NC.

Gopal, R., P. Goes, and R. Garfinkel (2002) Confidentiality via Camouflage: The CVC Approach to Database Security *Operations Research* **50** 501-516.

Jabine, Thomas B. (1993) Statistical disclosure limitation practices of United States statistical agencies. *Journal of Official Statistics* **9** 537-589.

Keller-McNulty, Sallie and Unger, Elizabeth A. (1993) Database systems: Inferential Security. *Journal of Official Statistics* **9** 475-499.

Kim, J. J. (1986) A method for limiting disclosure in microdata based on random noise and transformation. *Proceedings of the Survey Research Methods Section,* American Statistical Association, 370-374.

Kim, J. J., and Winkler, W. E. (1995), “Masking Microdata Files,”American Statistical Association, *Proceedings of the Section on Survey Research Methods*, 114-119.

Kooiman, P., Nobel, J. and Willenborg, L. (1999) Statistical data protection at Statistics Netherlands. *Special issue on statistical disclosure control. Netherlands Official Statistics* **14** 21-25.

Lambert, D. (1993) Measures of disclosure risk and harm. *Journal of Official Statistics* **9** 313-331.

Mackie, C. and Bradburn, N. (2000) Improving access to and confidentiality of research data. Washington, D.C.: National Academy Press.

Marsh, C., Skinner, C., Arber, S., Penhale, B., Openshaw, S., Hobcraft, J., Lievesley, D., and Walford, N. (1991) The case for samples of anonymized records from the 1991 census. *Journal of the Royal Statistical Society A* **154** 305-340.

Mood, A. M., Graybill, F. A., and Boes, D. C. (1963) *Introduction to the Theory of Statistics.* McGraw-Hill Press.

Moore, R. A. (1996), "Analysis of the Kim-Winkler Algorithm for Masking Microdata Files --- How Much Masking Is Necessary and Sufficient? Conjectures for the Development of a Controllable Algorithm," Statistical Research Division Report Series, **RR 95-05**, U.S. Bureau of the Census, Washington, DC.

Muralidhar, K., Batrah, D. and Kirs, P.J. (1995), “Accessibility, Security, and Accuracy in Statistical Databases: The Case for the Multiplicative Fixed Data Perturbation Approach,” *Management Science* 41(9). 1549-1584.

Muralidhar, K., Parsa, R. and Sarathy, R. (1999), “A General Additive Data Perturbation Method for Database Security,” *Management Science*, **45**(10), 1399-1415.

Muralidhar, K. and Sarathy, R. (1999) "Security of Random Data Perturbation Methods," *ACM Transactions on Database Systems*, **24**(4), 487-493.

Paass, G. (1988) Disclosure risk and disclosure avoidance for microdata. *Journal of Business and Economic Statistics* **6** 487-500.

Raghunathan, T., Reiter, J., and Rubin, D. (2003) Multiple Imputation for Statistical Disclosure Limitation. *Journal of Official Statistics,* **19**, 1-16.

Rubin, D. B. (1993) Discussion of statistical disclosure limitation. *Journal of Official Statistics* **9** 461-468.

Schlörer, J. (1981), "Security of Statistical Databases: Multidimensional Transformation," *ACM Transactions on Database Systems*, **6**, 91-112.

Spruill, N. L. (1983) Confidentiality and analytic usefulness of masked business microdata. The Public Research Institute. Alexandria, VA.

Sullivan, G. and Fuller, W. A. (1989) The use of measurement error to avoid disclosure. *Proceedings of the Section on Survey Research Methods.* American Statistical Association, 435-439.

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

Trottini, M. (2001) A decision-theoretic approach to data disclosure problems, *Research in Official Statistics,* **4,** 7-22.

Trottini, M. (2003) *Decision Models for Data Disclosure Limitation,* PhD thesis, Department of Statistics, Carnegie Mellon University, Pittsburgh.

Willenborg, L. and de Waal, T. (1996) *Statistical Disclosure Control in Practice*. Lecture Notes in Statistics **111** Springer, New York.

Zayatz, L., Moore, R. and Evans, B. T., (1996) New directions in disclosure limitation at the Census Bureau. U.S. Census Bureau Research Report No. LVZ96/01.