

February 7, 2003

Disclosure Risk vs. Data Utility: 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*

Recognizing that deidentification of data is generally inadequate to protect their confidentiality against attack by a data snooper, information organizations (IOs) can apply a variety of disclosure limitation (DL) techniques, such as topcoding, noise addition and data swapping. 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 coherent framework for examining tradeoffs between R and U for 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 of changes in the parameters of a DL procedure. Implementation of an R-U confidentiality map is illustrated in real multivariate data cases for two DL techniques: topcoding and multivariate noise addition. Topcoding is examined for a Cobb-Douglas regression model, as fit to restricted data from the New York City Housing and Vacancy Survey. Multivariate additive noise is examined under various scenarios of attack, predicated on different knowledge states for a data snooper, and for different goals of a data analyst. We illustrate how simulation methods can be used to implement an *empirical R-U confidentiality map*, which is suitable for analytically

intractable specifications of R, U and the disclosure limitation method. Application is made to the Schools and Staffing Survey, which is conducted by the National Center for Education Statistics.

(Additive Noise; Confidentiality; Disclosure Limitation; R-U Confidentiality Map; Topcoding)

1. Introduction: The Information Organization's Confidentiality Predicament

Statistical agencies and other information organizations (IOs) supply researchers and analysts with data obtained under confidentiality pledges from data providers (individuals, households and establishments). To assure confidentiality, an IO can lower the disclosure risk of a public-use data product by applying a disclosure limitation (DL) procedure to mask the data. Because this masking to lower disclosure risk will typically also lower the data utility, it is crucial that IOs assess the tradeoff. This article offers the *R-U confidentiality map* as an analytical framework for this assessment. For a general exploration of confidentiality and data access issues, see Duncan, Jabine and de Wolf 1993. The literature in disclosure limitation includes Duncan (2001), Fienberg (1994), Jabine (1993), Kooiman, Nobel and Willenborg (1999), Marsh *et al* (1991), Mackie and Bradburn (2000), Marsh, Dale and Skinner (1994) and Willenborg and de Waal (1996).

The IO's predicament is that data utility may suffer unduly or disclosure risk may yet remain too high with an improper choice of DL procedure or parameter value of a DL procedure. 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. In this article, we introduce the *R-*

U confidentiality map to provide such a foundation. 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 measured by the discrepancy between the masked data and the original data, it is called a *distortion measure* (Gomatam and Karr 2003). The R-U confidentiality map traces the joint impact on *R* and *U* of changes in parameter values of the DL procedure, enabling comparison of DL procedures and tradeoffs between disclosure risk and data utility.

To motivate the need for the R-U confidentiality map, here are three examples of information organizations' concern for disclosure limitation:

1. 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) add much to the data's utility but increase *confidentiality 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>)
2. Title 13, Section 9, of the U.S. Code requires that the U.S. Census Bureau disseminate no data product from which specific information about any particular respondent can be derived. Consistent with its mandate to provide data to a broad range of organizations, researchers and the public, the Census Bureau releases

microdata files that contain data from its censuses and surveys. They reduce the risk of disclosure using a variety of methods, including releasing only sample data, restricting geographic identifiers to areas of at least 100,000, top-coding, noise addition and data swapping.

3. Under authority for the National Center for Education Statistics (NCES), the National Education Statistics Act of 1994 prohibits these activities:
 - Using any individually identifiable information for any purpose other than statistical
 - Producing any publication in which data furnished by any particular individual can be identified
 - Permitting any person not authorized by the NCES Commissioner to examine any individual data or reports.

To comply with these statutory provisions, NCES applies various disclosure limitation methods. In public-use files developed from its School and Staffing Survey, for example, NCES “removes all state identifiers and stratum codes to prevent disclosing the identities of individual administrators and teachers. Detailed affiliation codes for private schools are collapsed into three categories: Catholic, Other Religious, and Nonsectarian. On the Administrator and Teacher files, income, age, and college or university attended are coded into categories.”

nces.ed.gov/surveys/SASS/confidentiality.asp

In all these cases, the 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?

The rudiments of the R-U confidentiality map were presented by Duncan and Fienberg (1999) and further explored for categorical data by Duncan *et al.* (2001). In this article, we show that the R-U map can be computed analytically in useful microdata cases, illustrated by the DL techniques of topcoding and multivariate noise addition. For intractable DL techniques, or complicated measures of R or U, we introduce the *empirical R-U confidentiality map*, a method suitable for a particular database. We illustrate its construction for the DL technique of topcoding, using data from the New York City Housing and Vacancy Survey, and for more complex implementations of multivariate noise addition, using data from the School and Staffing Survey of the National Center for Education Statistics.

2. Constructing an R-U Confidentiality Map: Topcoding

In its most basic form, an R-U confidentiality map is a set of paired values of disclosure risk R and data utility U, as generated by a family of DL strategies. To implement a DL procedure, the IO sets a parameter value, e.g., a topcoding threshold υ , that adjusts the stringency of disclosure limitation. The R-U confidentiality map shows the tradeoff between disclosure risk and data utility as a function of the parameter value. In this section, we illustrate for a topcoding strategy how an R-U confidentiality map can be developed for real bivariate data.

For the 1999 New York City Housing and Vacancy Survey (www.census.gov/hhes/www/housing/nychvs/abstract.html), the Census Bureau determined that 1/2 of 1 percent of the renter-occupied units surveyed had a contract rent above \$2950. The released microdata about rental unit characteristics and their occupants contained the monthly rent variable. However the 81 units with rent above $\upsilon = \$2950$ were topcoded and reported as

having a rent of \$3817, which is the (conditional) mean rent for these cases. We develop an R-U confidentiality map to examine the impact of this topcoding procedure and its choice of threshold on disclosure risk and data utility. As υ is changed, a curve is mapped in the R-U plane, portraying the tradeoff between disclosure risk and data utility as υ is lowered and more extensive masking imposed.

A data analyst examines the relationship between rent and household income for households with income above \$1000. Using an elementary Cobb-Douglas model, and ignoring the fact that the data have been topcoded, the analyst estimates the regression equation for log rent on log income as $\ln \text{rent} = 3.320 + 0.3039 (\ln \text{income})$. The standard error of regression is $S=0.5287$, standard error of the regression coefficient of log income is 0.005391, and coefficient of determination is $R^2 = 24.1 \%$.

To assess data utility, we compare these results with those obtained if there had been no topcoding. While without the actual confidential data this cannot be determined, for the sake of this illustration, we imputed the 81 topcoded rent values. With these data, we obtain regression results of $\log \text{rent} = 3.323 + 0.3035 \log \text{income}$, with standard error of the coefficient of log income 0.005392 and coefficient of determination $R^2 = 24.1 \%$. The data utility of the topcoded data can be assessed in a variety of ways. Since the estimated coefficient of the regressor is typically of interest, take U to be the reciprocal of the squared difference between the estimated regression coefficient of log income from the original and imputed data and the regression coefficient of log income from the topcoded data set. Divided by 10^7 for convenience in scaling, this yields $U = 0.625$.

Modeling disclosure risk requires assumptions about the goals of the data snooper. Duncan and Lambert (1989) lay out two types of confidentiality disclosure: identity disclosure

(Is the data snooper able to correctly link a record in the released data to a known individual?) and attribute disclosure (How close is the data snooper to determining an attribute's value for an individual?). Note that identity disclosure either happens or not, while attribute disclosure is a matter of degree. While R can be structured to measure either form of disclosure risk in this example, take the concern to be identity disclosure and define R to be the *maximum* probability of disclosure over all individuals in the data. In this example, assume a logistic model between π = the probability of identification of a record and x = its rental value; i.e.,

take $\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta x$. As can be done in practice even without applicable empirical

evidence, we obtain the parameters of this model by assessing two points using judgment. In this case, we assess the probability π that a data snooper can identify a record when the released rental value x is \$500 to be 0.01, and the probability π of identification for a released rental value x of \$5000 to be 0.5. (In practice, an IO would examine the sensitivity of the resulting R-U confidentiality map to this specification, but for this illustration we will leave it at that.)

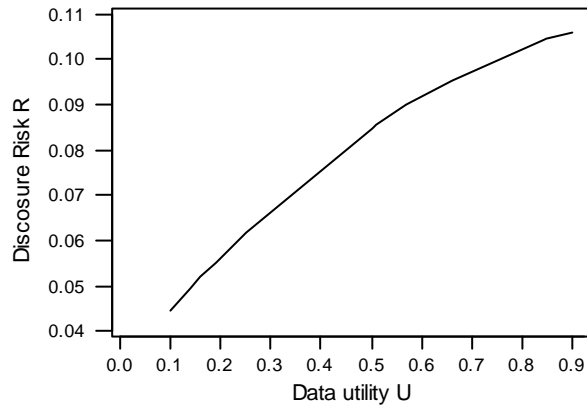
Reasonably, renters with rental values that have been topcoded are taken to have negligible probability of disclosure when only the average value of \$3817 is released for them. Since the logistic model has π increasing in x , R is then the value of π given by the logistic model at the topcoding threshold v . For the threshold used by the Census Bureau (\$2950), we compute $R = 0.11$. Thus for this implementation, the R-U value is (disclosure risk, data utility) = (0.11, 0.625).

Therefore, according to the model fit to the topcoded data, a doubling of income predicts a 23.45% increase in rent, while according to the model fit to the original (and imputed) data, a doubling predicts a 23.41% increase in rent, a minimal difference. In this case, the data utility is adequate for any practical purpose. If a disclosure risk of 0.11 is considered excessive, the IO

could examine the tradeoff between R and U as the topcoding threshold is lowered by constructing the R-U confidentiality map.

To obtain the R-U confidentiality map, repeat the above process for a variety of values of the topcoding threshold ν , ranging from \$2950 to \$1800. Fitting a smoothed curve to the resulting points gives Figure 1. Using this R-U confidentiality map, the IO can determine that to obtain a disclosure risk below 0.08, say, requires a topcoding threshold value of $\nu = 2560$. In that

Figure 1. R-U Confidentiality Map: Topcoding



case the data utility is about 0.5.

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

In this section, we construct an R-U confidentiality map for multivariate additive noise (Spruill 1983, Paass 1988, Sullivan and Fuller 1989, Duncan and Mukherjee 2000), using an implementation discussed by Kim (1986). Through calculations based on theoretical distributions and under a variety of realistic assumptions about the state of knowledge and motivation of the data snooper, we demonstrate that the assessment of disclosure limitation methods depends on data snooper behavior. Take the original data to be $\mathbf{X} = [X_{ij}] = [\mathbf{X}_1, \mathbf{K}, \mathbf{X}_n]'$,

where $\mathbf{X}_i : (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and the masked data have the additive noise form, $\mathbf{Y} = \mathbf{X} + \boldsymbol{\varepsilon}$, where

$$\boldsymbol{\varepsilon}_{n \times p} \sim (\mathbf{0}, \lambda^2 \boldsymbol{\Sigma}).$$

In practice, the form of R and U should be tailored to the particular situation at hand. For purpose of illustration, take the data utility U to be the reciprocal of the data user's mean squared error in estimating linear combinations $\mathbf{c}'\boldsymbol{\mu}$ of the components of the population mean vector $\boldsymbol{\mu}$, and the disclosure risk R to be the reciprocal of the mean squared error the data snooper can achieve in inferring the value of a target value τ for an individual entity. To quantify disclosure risk R, consider three different knowledge states, depending on the group to which the data snooper can isolate the target τ :

1. *Population.* τ has the same distribution as one of the attributes \mathbf{X}_j
2. *Sample.* τ is one of the values in \mathbf{X} , (i.e., is in the sample)
3. *Record.* τ is not only in the sample, but the data snooper has enough external information to be able to identify (link to) the specific record to which the target belongs in \mathbf{X} .

The first case is appropriate when the data are a small sampling fraction from a population and the data snooper cannot be sure that the target entity is in the sample. The second case is appropriate when the data are a census or a near census. The third case is appropriate when the data snooper has external identifying information that permits linkage to the target record. Results will be developed and discussed for all three states of knowledge and under various goals of the data snooper, i.e., to compromise a specific entity or a fishing expedition for any entity. In Section 3.1, we examine the three states of knowledge above and the data snooper takes the target τ to be typical of the data. In Section 3.2, we examine parallel states of

knowledge where the data snooper knows in addition that the target τ is at a certain percentile point. In Section 3.3, we instead take the data snooper to know that the target τ is an extreme value.

3.1. Target Typical of Data: Three Knowledge States

Data Utility: The data user estimates the population mean vector $\boldsymbol{\mu}$ using $\hat{\boldsymbol{\mu}} = \bar{\mathbf{Y}}$, the sample

mean of the masked data. Therefore, $E(\hat{\boldsymbol{\mu}}) = \boldsymbol{\mu}$ and $Var(\hat{\boldsymbol{\mu}}) = \frac{1 + \lambda^2}{n} \boldsymbol{\Sigma}$. For a goal of estimating an

arbitrary linear combination $\mathbf{c}'\boldsymbol{\mu}$, take the data utility to be the reciprocal of the mean squared

error and so $U = \frac{n}{1 + \lambda^2} (\mathbf{c}'\boldsymbol{\Sigma}\mathbf{c})^{-1}$. Note that this data utility measure takes into account the

multivariate, correlational structure of the data.

Disclosure Risk: With the data snooper's goal to compromise a specific entity, the first two states of knowledge, *Population* and *Sample*, yield the same disclosure risk. With the data snooper having a specific target value τ in attribute j and using $\hat{\tau} = \bar{Y}_j$, the disclosure risk is

$$R = \frac{1}{E(\tau - \hat{\tau})^2} = \frac{n}{(1 + \lambda^2)\sigma_j^2 + n(\mu_j - \tau)^2}. \quad (1)$$

The third state of knowledge, *Record*, in which the data snooper is able to identify the masked record that corresponds exactly to the target τ , confronts the IO with the worst disclosure risk. Here, if the snooper uses $\hat{\tau} = Y_{ij} = \tau + \varepsilon_{ij}$, the risk is

$$R = \frac{1}{E(\tau - \hat{\tau})^2} = \frac{1}{E(\varepsilon_{ij})^2} = \frac{1}{\lambda^2\sigma_j^2}. \quad (2)$$

Whatever the knowledge state of the data snooper, the disclosure risk without data masking is found by setting $\lambda^2 = 0$. Note that under the third state of knowledge, R is infinite

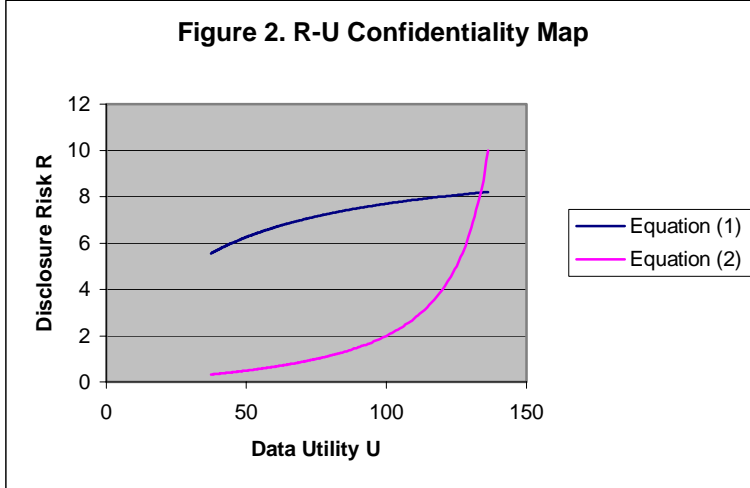
without masking. Thus, in circumstances where record linkage is feasible, release of the original data would pose too much of a threat to confidentiality.

Is the data snooper always better off—when knowing the target’s index i —using $\hat{\tau} = Y_{ij}$ to assess the target value $t = x_{ij}$? Comparing Equations (1) and (2), the data snooper actually

gains by using \bar{Y}_j , whenever $\lambda^2 > \left(\frac{n}{n-1}\right)\left(\frac{\tau - \mu_j}{\sigma_j}\right)^2 + \frac{1}{n-1}$, which for large n is approximately

the square of the number of standard units the target τ lies from the mean. Thus, by adding sufficient noise, the IO can eliminate the advantage the data snooper has through record linkage.

Displayed in Figure 2 is an R-U confidentiality map for the two risk measures in this example (with $n = 50$, $\sigma_j^2 = 1$, $\mathbf{c}'\Sigma\mathbf{c} = \mathbf{3}$, and $(\mu_j - \tau)^2 = 0.1$). The figure shows the impact on data utility and disclosure risk of changes in the disclosure limitation parameter λ^2 , when the data snooper knows the index of the target (Equation (2)) and when the data snooper does not (Equation (1)). With additive noise masking, knowing the index does not help the data snooper once the extent of noise is large enough. As Figure 2 shows, when the curve for Equation (2) crosses below that for Equation (1), the data snooper is better off ignoring knowledge of the index of the target. If so, the disclosure risk would switch to the upper curve in this domain of larger λ^2 .



3.2 Target at a Percentile Point

Extreme or outlying data values may present targets that are easily compromised through record linkage. For the IO, this vulnerability is doubly serious because disclosure of targets with atypical values may pose more serious consequences; for example they may be for high profile respondents. For such targets, the data snooper can use a specified percentile point or an extreme value for an estimator of the target's attribute value. In this section, we explore the percentile case and, in the next section, we explore the extreme value case.

For the percentile case, consider two states of knowledge for the data snooper:

1. *Population knowledge.* The target τ is the p^{th} *population* percentile point
2. *Record knowledge.* The target τ is known to be in the sample and to be the p^{th} *sample* percentile point.

We establish notation and structure: For the attribute of the target τ , let the data be generated as $X_1, X_2, \dots, X_n \sim iid(\mu, \sigma^2)$, with the realizations, x_1, \dots, x_n . Denote the p^{th}

population percentile point as $\xi_{xp} = \mu + l_{xp}\sigma$, so that $F_X(\xi_{xp}) = P(X \leq \mu + l_{xp}\sigma) = p$. Denote the released values for the attribute of the target as $Y_i = x_i + \varepsilon_i$, $\varepsilon_i \sim iid(0, \lambda^2\sigma^2)$, $i = 1, \dots, n$. Finally, denote the p^{th} population percentile point of the masked data as $\xi_{yp} = \mu + l_{yp}\sqrt{\sigma^2(1 + \lambda^2)}$, so that $F_Y(\xi_{yp}) = P(Y \leq \mu + l_{yp}\sqrt{\sigma^2(1 + \lambda^2)}) = p$.

Consider a data snooper attack through the obvious estimator $\hat{\tau} = \hat{\xi}_{yp} = Y_{(np)}$, for either the population or sample percentile, depending on population knowledge or record knowledge. This estimator is biased when the data have been altered by additive noise. Consider a value of p so that np is an integer. With strict monotonicity of the distribution function $F_Y(\cdot)$, the percentile estimator has the asymptotic distribution (see Mood, Graybill and Boes, 1963, p. 257),

$$\hat{\tau} = Y_{(np)} \sim N\left(\xi_{yp}, \frac{p(1-p)}{n[f_Y(\xi_{yp})]^2}\right).$$

Population knowledge. With the data snooper knowing that τ is the p^{th} population percentile, the disclosure risk is approximately given by

$$R = \frac{1}{E(\tau - \hat{\tau})^2} \approx \frac{1}{\frac{p(1-p)}{n[f_Y(\xi_{yp})]^2} + \left(l_{yp}\sqrt{\sigma^2(1 + \lambda^2)} - l_{xp}\sigma\right)^2}. \quad (3)$$

Assuming X_i and ε_i are normally distributed, $l_{xp} = l_{yp} = z_p$, and $f_x(\xi_{xp}) = \frac{\varphi(z_p)}{\sigma}$

and $f_y(\xi_{yp}) = \frac{\varphi(z_p)}{\sqrt{\sigma^2(1 + \lambda^2)}}$, where z_p is the p^{th} percentile point of a standard normal distribution

and $\varphi(\cdot)$ is the standard normal density function. In that case, Equation (3) becomes

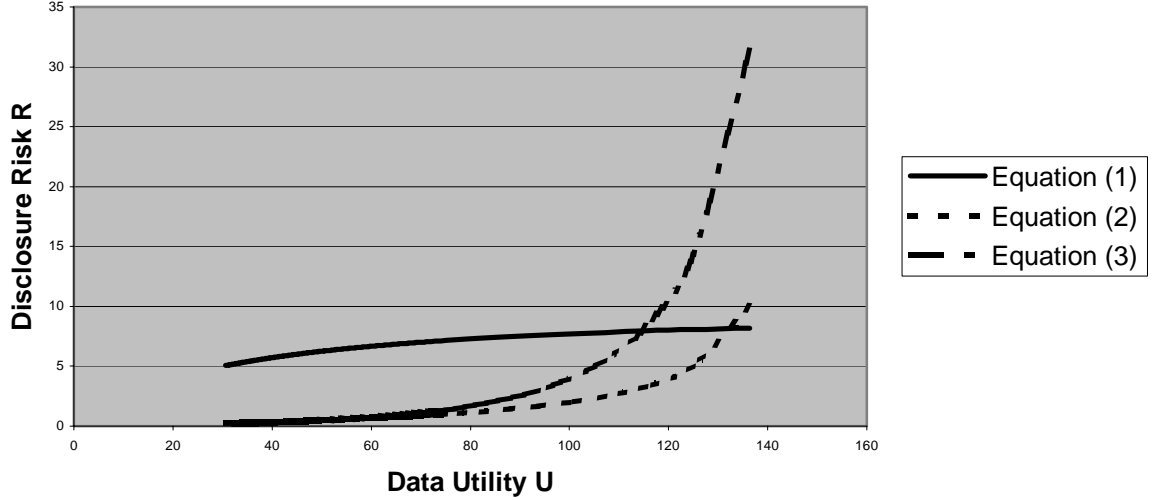
$$R \approx \frac{1}{\left(\frac{p(1-p)}{n[\varphi(z_p)]^2} \right) \sigma^2 (1 + \lambda^2) + z_p^2 \left(\sqrt{\sigma^2 (1 + \lambda^2)} - \sigma \right)^2}. \quad (4)$$

By Equation (4), the IO can lower its disclosure risk adequately by setting λ^2 sufficiently large. The IO can then anticipate that the data snooper will be deterred from making an attribution for the value of τ based on the estimator, $\hat{\xi}_{y_p}$. Alternatively, however, the data snooper might employ a different estimator. Alternative estimators could be based on the data snooper either knowing the value of λ or not. If the data snooper knew the value of λ , the data snooper could make an adjustment for the bias in $\hat{\xi}_{y_p}$. The data snooper has two possible sources of information about the value of λ : (1) *revelation*—the IO could have revealed the value it used in masking, or (2) *experience*—based on experience with similar data, the data snooper may have a strong prior belief about σ and so can back out an estimate of λ from the sample variance of the masked data. Because of (1), the IO has to realize that revelation—publicly releasing its value of λ —could have a detrimental effect on disclosure risk (although it may also have a positive effect on data utility). If circumstances require concern for the experience of the data snooper, as in (2), the IO may need to consider disclosure limitation methods other than additive noise. Without knowing the value of λ , the data snooper might consider \bar{Y} as an alternative estimator of the target value τ , essentially admitting that the IO's use of additive noise has made inoperative the data snooper's use of knowledge that the target is at a particular percentile point. Comparing Equations (1) and (3) gives circumstances when the data snooper should switch to \bar{Y} . Displayed as Figure 3 is an R-U confidentiality map comparing the three knowledge states of the data snooper as they affect disclosure risk given by Equations (1), (2) and (3). Parameter values $n = 50$, $\sigma^2 = 1$, and $(\mu - \tau)^2 = .1$, and $p = .95$ are used. Note that the data snooper should switch to \bar{Y}

whenever the curve for Equation (1) is above the curve for Equation (3), so when $\lambda^2 \geq 0.17$.

Thus, the advantage to the data snooper of knowing the population percentile point of the target is obviated when the amount of additive noise is this large.

Figure 3. R-U Confidentiality Map



Record knowledge. The data snooper knows that the target is the *sample* p^{th} percentile point, so $\tau = x_{(np)}$. Still using $\hat{\tau} = \hat{\xi}_{Y_p} = Y_{(np)}$, the disclosure risk is

$$R = \frac{1}{E(\tau - \hat{\tau})^2} = \frac{1}{\text{Var}(\hat{\tau}) + (\tau - \xi_{Y_p})^2}. \quad (5)$$

Equation (5) yields no insight into how disclosure risk behaves over the range of p .

Restating Equation (5) as

$$R = \frac{1}{\sum_{i=1}^n E_{\varepsilon} [P(\hat{\tau} = x_i + \varepsilon_i) (\tau - x_i - \varepsilon_i)^2]}, \quad (6)$$

we see that the disclosure risk R is maximized when the np^{th} masked value corresponds exactly

to the np^{th} original data point, or $\hat{\tau} = x_{(np)} + \varepsilon_{(np)}$. In this case, $R = \frac{1}{\lambda^2 \sigma_j^2}$, which is identical to

Equation (2), where the data snooper is able to link exactly to the target's masked value. This is most likely to be true with a target in the extremes (e.g., large p) because data tend to spread apart more in the tails of common unbounded distributions. Thus, the IO would need large values of λ^2 to misalign the ordering of the extremes for the masked sample versus the unmasked sample.

3.3 Target at an Extreme

Suppose the data snooper knows that a target is one of the extreme values in the released data: it is either the largest or smallest in a file. As in the previous section, the data have been masked using additive noise. If the data snooper knows that the target is the maximum in the sample, a natural attack estimator would be $\hat{\tau} = Y_{(n)}$. The risk is then the reciprocal of

$E(\tau - \hat{\tau})^2 = E(x_{(n)} - Y_{(n)})^2 = \text{Var}(Y_{(n)}) + (x_{(n)} - \mu_{Y_{(n)}})^2$, where $\mu_{Y_{(n)}}$ is the mean of the maximum order statistic $Y_{(n)}$. For a normally distributed attribute, this expression can be evaluated for any finite n by using a table of moments of normal order statistics. For large samples, the classic results of Fisher and Tippett (1928) show that $\mu_{Y_{(n)}} \approx \mu + \sqrt{\sigma^2(1 + \lambda^2)}K_{n1}$, where

$$K_{n1} = \sqrt{2 \log n} - \frac{\log \log n + \log 4\pi - 2\gamma}{2\sqrt{2 \log n}} \text{ and } \gamma = 0.57722 \text{ is Euler's constant; and}$$

$$\text{Var}(Y_{(n)}) \approx \sigma^2(1 + \lambda^2)K_{n2}, \text{ where } K_{n2} = \frac{\pi^2}{12 \log n}. \text{ [An accessible reference for these results is}$$

Cramér (1946), p. 376.] Thus,

$$R = \frac{1}{E(x_{(n)} - Y_{(n)})^2} \approx \frac{1}{K_{n2}\sigma^2(1 + \lambda^2) + (x_{(n)} - (\mu + K_{n1}\sigma^2(\sqrt{1 + \lambda^2} - 1)))^2}. \quad (7)$$

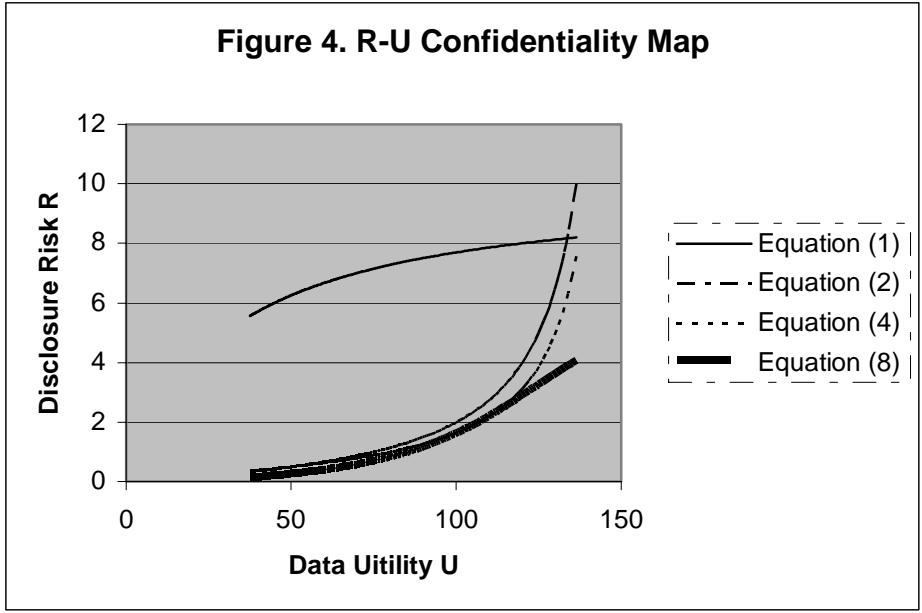
Similar results can be derived for other extreme values, such as the minimum, or even the r^{th} largest or smallest for small r.

To demonstrate the R-U confidentiality map for this example substitute $x_{(n)} \approx \mu + \sigma K_{n1}$ into Equation (7), to get a typical value for the extreme. This gives a risk of

$$R \approx \frac{1}{K_{n2}\sigma^2(1 + \lambda^2) + K_{n1}^2\sigma^2(\sqrt{(1 + \lambda^2)} - 1)^2} \quad (8)$$

Note the similarity between Equations (8) and (4): both denominators are linear combinations of $\sigma^2(1 + \lambda^2)$, the variance of a masked observation, and $\sigma^2(\sqrt{1 + \lambda^2} - 1)^2$, which measures the discrepancy between the standard deviation of a masked observation and the standard deviation of an original observation. Note also that for large samples, the disclosure risk given by Equation (8)—so the target is an extreme—goes to zero, while the disclosure risk given by Equation (4)—so the target is a percentile point—goes to a positive value.

Displayed in Figure 4 is an R-U confidentiality map for the impact on data utility and disclosure risk of the disclosure limitation parameter λ^2 , under the assumption that the data snooper knows the target is the maximum in the sample (Equation (8)), knows the 99th population percentile is the target (Equation (4)), only knows the target is from the same population as the sample (Equation (1)), and knows how to link the masked value to the target (Equation (2)). Parameter values $n = 50$, $\sigma^2 = 1$, and $(\mu - \tau)^2 = .1$, and $p = .95$ are used. Note that knowing that the target is an extreme benefits the data snooper less than knowing that the target is at a percentile point or knowing the index of the target.



4. A Database-Specific Approach: Constructing an Empirical R-U Confidentiality Map

In the previous sections we developed theory for the R-U confidentiality map and showed insights for disclosure limitation procedures. In this section we formally introduce the *empirical R-U confidentiality map*. Like the topcoding example of Section 2, it is based on a particular database, and can be constructed even if the data, masking method, or snooper strategy do not allow tractable theoretical analysis. Using real-life data of both practical size and realistic complexity, we detail how this empirical R-U confidentiality map can be used to:

- Examine the impact on disclosure risk and data utility of various types of data snooper knowledge
- Facilitate comparisons between various disclosure limitation methods.

To illustrate the empirical R-U confidentiality map, we use data from the Teacher Followup Survey (TFS) of 1994-95 by the National Center for Education Statistics (NCES). This

survey involved a sample of teachers first interviewed in 1993-94 under the School and Staffing Survey (SASS). The goal of the TFS was to provide data for an investigation of attrition rates for teachers, and to elicit characteristics and attitudes of leavers and non-leavers from the profession. For that survey data on the nation’s part-time private school teachers, we examine total earned income from all sources (identified as TFS376 in the survey documentation) and base salary (TFS368). Summary statistics for the two attributes are shown in Table 1 (the value of the sample size n is considered confidential by NCES).

Table 1. Summary Statistics for Income and Salary in the 1993-94 Teacher Followup Survey. Part-Time Private School Teachers.

Attribute	n	Sample Mean	Sample SD	Min	1 st %-ile	10 th %-ile	90 th %-ile	99 th %-ile	Max
1: Income	na	\$20.1K	\$13.3K	\$2K	\$2K	\$7.2K	\$35K	\$70K	\$95.2K
2: Salary	na	\$15.5K	\$10.2K	\$0	\$0	\$4.5K	\$28K	\$55K	\$65.0K

Our analysis concerns estimation of two parameters of this population: (1) the difference in means of total income and base salary, $\mu_1 - \mu_2$, and (2) the regression coefficient for the simple regression of income on salary, β_1 . The data snooper wants to infer the income of a particular target record τ based on a specific type of knowledge. We consider targets of maximum, minimum, and p^{th} sample percentile ($p = 0.01, 0.10, 0.90, \text{ and } 0.99$). The data snooper has two possible types of record knowledge:

1. *Index knowledge:* $\hat{\tau} = Y_{1t}$, where t is the index of the record in the unmasked sample which corresponds to the largest, smallest, or p^{th} sample percentile, or
2. *Position knowledge:* $\hat{\tau} = \max_i(Y_{1i}), \min_i(Y_{1i}), \text{ or } Y_{1(np)}$

Under index knowledge, the data snooper knows the index of the target; it is the same as the case where the data snooper has enough external information to identify (link to) a specific masked

record. Under position knowledge, the data snooper knows the position of the target in the unmasked data.

We study a modified version of noise addition. Specifically, since income and salary are bounded below by 0, their masked values are truncated at 0, so income and salary for the i^{th} record, x_{1i} and x_{2i} appear in the released file as

$$Y_{1i} = \max(0, x_{1i} + \varepsilon_{1i}) \text{ and } Y_{2i} = \max(0, x_{2i} + \varepsilon_{2i}), \quad (9)$$

where $(\varepsilon_{1i}, \varepsilon_{2i})$ are bivariate normal random vectors with mean $\mathbf{0}$ and covariance matrix $\lambda^2 \hat{\Sigma}$,

where $\hat{\Sigma}$ is the estimated variance covariance matrix of the unmasked variables x_{1i} and x_{2i} .

Because the masking in Section 3 did not include truncation, the expressions for data utility and disclosure risk presented there are not directly applicable.

To obtain the empirical R-U confidentiality map, first simulate the masking process for a range of values of the masking parameters by generating a number, say M , of masked datasets.

From these simulated datasets, the disclosure risk $R = 1/E(\tau - \hat{\tau})^2$ and the data utility $U =$

$1/E(\hat{\theta}_y - \theta_x)^2$ are estimated, where θ_x is the data analyst's parameter of interest and $\hat{\theta}_y$ is the

estimator of that parameter computed from the masked data. Estimate R for each λ^2 from the M

simulated datasets by

$$\left(\frac{1}{M} \sum_{m=1}^M (\tau - \tau_m)^2 \right)^{-1}, \quad (10)$$

where $\hat{\tau}_m$ is the data snooper's prediction of the target in replicate m . To estimate U , express the

mean squared error as

$$E(\hat{\theta}_y - \theta_x)^2 = \text{Var}(\hat{\theta}_y) + [E(\hat{\theta}_y) - \theta_x]^2 \quad (11)$$

To estimate the variance of the right hand side of (11) for any value of λ^2 , use the average of the estimates of $Var(\hat{\theta}_y)$ as computed from the simulated datasets:

$$\hat{V}ar(\hat{\theta}_y) = \frac{1}{M} \sum_{m=1}^M \hat{v}(\hat{\theta}_{ym}). \quad (12)$$

Similarly, the squared bias term on the right hand side of (11) is estimated as

$$\hat{B}^2(\hat{\theta}_y) = \left(\frac{1}{M} \sum_{m=1}^M \hat{\theta}_{ym} - \hat{\theta}_x \right)^2, \quad (13)$$

where $\hat{\theta}_{ym}$ is the estimate of θ_x made from the m th masked dataset. Then U is estimated as the reciprocal of the sum of (12) and (13).

We carry out this procedure using $M = 200$ and values of λ^2 ranging from 0 to 1.00. We calculate estimates of $\mu_1 - \mu_2$ and β_1 and their variances for each replicate and each λ^2 . For the difference in means, we calculate $\hat{\theta}_{ym} = \bar{Y}_{1m} - \bar{Y}_{2m}$ and $\hat{v}(\hat{\theta}_{ym}) = (s_{y_1m}^2 + s_{y_2m}^2 - 2s_{y_1y_2m})/n$, where $s_{y_1m}^2$, $s_{y_2m}^2$, and $s_{y_1y_2}$ are the sample variances and covariance of the two variables from the m^{th} masked dataset. Then from (11) – (13), an estimate of the utility U is

$$\left[\frac{1}{M} \sum_{m=1}^M \mathbf{c}' \hat{\Sigma}_{ym} \mathbf{c} / n + \left(\left[\frac{1}{M} \sum_{m=1}^M (\bar{Y}_{1m} - \bar{Y}_{2m}) \right] - (\bar{x}_1 - \bar{x}_2) \right)^2 \right]^{-1}. \quad (14)$$

Similarly, for the regression parameter, calculate $\hat{\theta}_{ym} = \hat{\beta}_1 = \frac{s_{y_1y_2m}}{s_{y_2m}^2}$ and $\hat{v}(\hat{\theta}_{ym}) = \hat{v}(\hat{\beta}_{1m}) = \frac{\hat{\sigma}_{em}^2}{s_{y_2m}^2}$,

where $\hat{\sigma}_{em}^2$ is the estimated error variance from the regression on the m^{th} masked dataset.

Figure 5 shows the empirical R-U confidentiality map for the case in which the data snooper knows the target to be the maximum or the minimum in the sample and the analyst is estimating the difference in means. The risk falls rapidly as λ^2 increases from 0, but then levels

off with little additional protection from disclosure though utility suffers severely. Note that the disclosure risk remains high for all values of λ^2 when the data snooper's target is the minimum. This is because truncation of the masked income at 0 limits the magnitude of the downside error to \$2K, which has a high associated risk value. The empirical R-U confidentiality map for the regression parameter β_1 has a similar form, so is not shown here.

For analyst interest in the regression parameter β_1 , Figure 6 is an empirical R-U confidentiality map that compares the disclosure risk for two different knowledge states for the data snooper about the same record: the target is the 99th sample percentile (\$70K for this sample) and the data snooper can link the masked value to this target. This map shows that the disclosure risk with data snooper knowledge of the percentile is greater than that with knowledge of the index of the record. Thus, this example substantiates that an appropriate masking strategy should depend on what knowledge the data snooper has about the target.

Figure 5. Empirical R-U Confidentiality Map: Comparing Risk When the Data Snooper Knows the Targeted Record to be a Maximum or a Minimum and Analyst Estimates Mean Difference

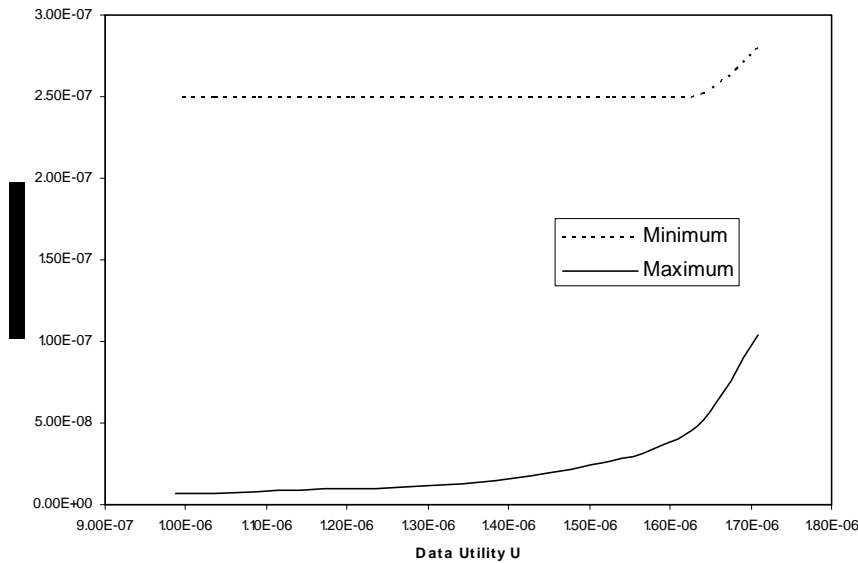
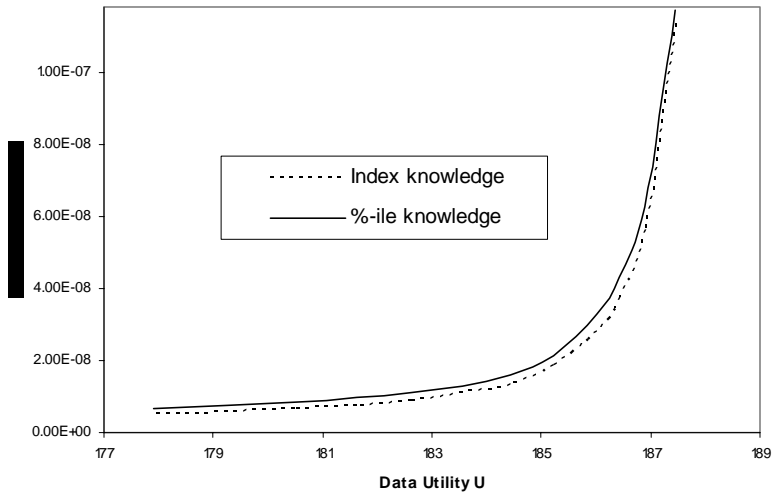


Figure 6. Empirical R-U Confidentiality Map: Comparing Risk When the Data Snooper Knows the Index or the Percentile of the Targeted Record (at the 99th percentile) and the Analyst Estimates Regression Coefficient



Finally, Figures 7a and b show how an empirical R-U confidentiality map can help compare two alternative masking methods. “Equal variance masking” is implemented as in Equation (9). In “unequal variance masking”, the variance of the noise is doubled for sensitive data, which in this case is defined as an income value exceeding \$35K. Figure 7a shows the R-U confidentiality map for the utilities of the estimator of the difference in means for these two masking methods, while Figure 7b shows the map for the regression coefficient estimator. In both cases, the data snooper uses knowledge of the percentile to identify the target value. The figures show that the analyst’s parameter of interest determines which disclosure limitation method is better for the IO. Unequal variance masking will allow greater protection against the data snooper than equal variance masking while maintaining utility for the difference in means. However, the reverse is true for estimating the regression coefficient, since the extreme values provide the most valuable information for its estimation. Of course, in practice an IO must be prepared to protect against a variety of snooper targets and strategies, and to provide high utility for many potential estimators. Thus the decision about the type of masking to use should not be

made on the basis of a single R-U confidentiality map. Nonetheless, such maps do provide the basis for addressing the problem. Indeed, in practice a weighted average, that reflects assessments of conditions of attack and usage, of R-U confidentiality maps may well be of considerable value.

Figure 7a. Empirical R-U Confidentiality Map: Comparing Risk for Two Masking Methods When Target is Known to be at 99th %-ile and Analyst Estimates Mean Difference

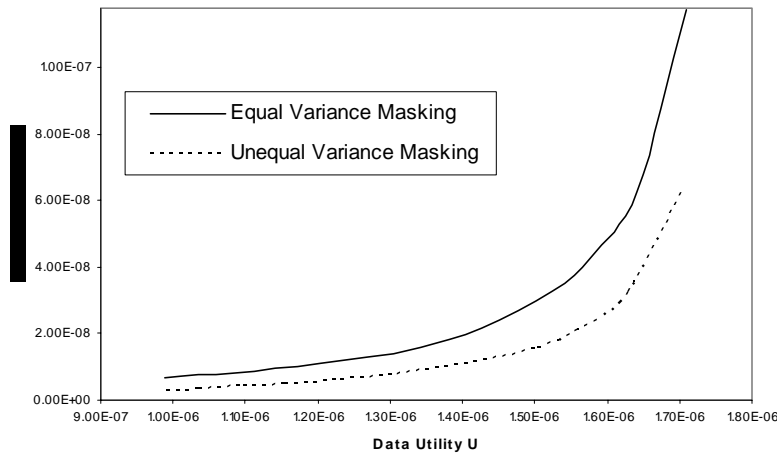
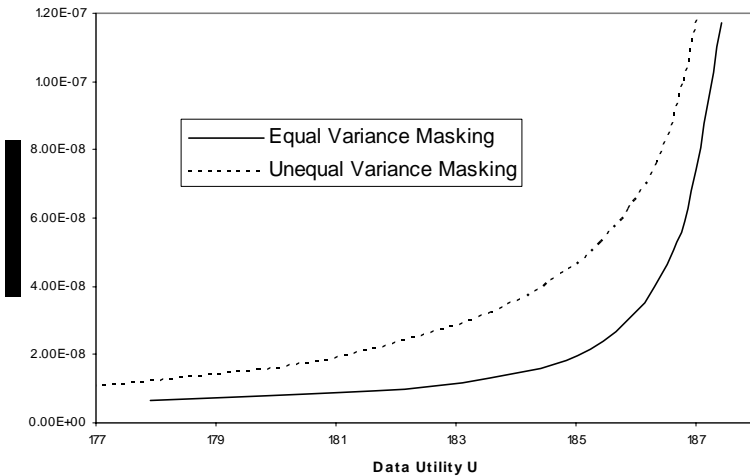


Figure 7b. Empirical R-U Confidentiality Map: Comparing Risk for Two Masking Methods When Target is Known to be at 99th %-ile and Analyst Estimates Regression Coefficient



5. Conclusions

This article develops a tool that can help IO's make better decisions about disclosure limitation methods that can fulfill their dual mandate of providing useful data while maintaining an adequate level of confidentiality. We studied the tool for two DL methods, topcoding and multivariate noise addition, and implemented it on real data sets to provide examples of specific approaches that can be used to model R and U. In the topcoding case, the analyst estimated a Cobb-Douglas regression coefficient. In the multivariate noise addition case, the analyst estimated the difference of two means and a regression coefficient. A notable contribution is a demonstration of how R-U confidentiality maps can compare (1) different states of data snooper knowledge and (2) different disclosure limitation methods.

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 become more 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 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 microaggregation and binning (Domingo-Ferrer and Torra 2001).

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, Karthik Kannon and Laura Zayatz. The authors also thank the National Center for Education Statistics for providing—under nondisclosure license 010207550—access to the individually identifiable survey database entitled, “National Education Longitudinal Study of 1988 (NELS), Schools and Staffing Survey (SASS), and all follow-ups.”

Our thanks go also to two referees and the associate editor for helpful comments and suggestions.

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.

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

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. 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.
- 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.
- S. Gomatam and A. F. Karr (2003). Distortion measures for categorical data swapping. Technical Report, National Institute of Statistical Sciences. Research Triangle Park, NC.
- Jabine, Thomas B. (1993) Statistical disclosure limitation practices of United States statistical agencies. *Journal of Official Statistics* **9** 537-589.
- 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.
- 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., Dale, A. and Skinner, C. J. (1994) Safe data versus safe settings: Access to microdata from the British Census. *International Statistical Review* **62** 35-53.
- 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.
- Muller, W., Blien, U. and Wirth, H. (1995) Identification risks of microdata. *Sociological Methods and Research* **24** 131-157., G. and Wasuchkuhn, U. (1985) Datenzugang, datenschutz, und anonymisierung. *Analysepotential und Indentifizierbarkeit von Anonymisierten Individualdaten*. Munich and Vienna: R. Oldenbourg.
- Paass, G. (1988) Disclosure risk and disclosure avoidance for microdata. *Journal of Business and Economic Statistics* **6** 487-500.
- Rubin, D. B. (1993) Discussion of statistical disclosure limitation. *Journal of Official Statistics* **9** 461-468.
- 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.

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