NISS

Disclosure Risk vs. Data Utility: The R-U Confidentiality Map

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

Technical Report Number 121
December, 2001

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

Also, a Los Alamos National Laboratory Technical Report, LA-UR-01-6428.

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

Information organizations (IOs) must provide data products that are both useful and have low risk of confidentiality disclosure. Recognizing that deidentification of data is generally inadequate to protect their confidentiality against attack by a data snooper, concerned IOs can apply disclosure limitation techniques to the original data. Desirably, the resulting restricted data have both high data utility U to users (analytically valid data) and low disclosure risk R (safe data). This article shows the promise of the *R-U confidentiality map*, a chart that traces the impact on R and U of changes in the parameters of a disclosure limitation procedure. Theory for the R-U confidentiality map is developed for additive noise applied to univariate data under various scenarios of data snooper attack. These scenarios are predicated on different knowledge states for the data snooper. A demonstration is provided of how to implement the theory for a real database. Through simulation methods, this leads to an *empirical R-U confidentiality map*. Application is made to data from a National Center for Education Statistics (NCES) survey, the Schools and Staffing Survey (SASS).

(*Additive Noise; Confidentiality; Disclosure Limitation; Data Snooper; Data Utility; Disclosure Risk; Record-Specific Risk; Restricted Data; R-U Confidentiality Map*)

1. Introduction: The Information Organization's Confidentiality Problem

Empirical analysis requires access to data. For data about important policy and management issues, information organizations (IOs)—such as government statistical agencies—are the conduit between data providers and data users. Potentially blocking flow is the fact that data collected by an IO from respondents (units of data collection, e.g., individuals, households, establishments, etc.) are subject to pledges of confidentiality (Marsh, Dale and Skinner 1994). Confidentiality arises both from ethical concerns about the autonomy of the respondent and pragmatic concerns about quality and quantity of response. Pledges may be either implicit or explicit, and in some cases are specified by regulation or legal statute, such as the European Union's Data Protection Directive. (For the general issues, see Duncan, Jabine and de Wolf 1993).

Here are some examples of how confidentiality is a concern for a variety of information organizations:

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."
2. Title 13 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.
3. 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, or

- Permitting any person not authorized by the NCES Commissioner to examine any individual data or reports.

4. The HIPAA Privacy Rule took effect on April 14, 2001. Within two years from that date, this regulation obligates most covered entities (health plans, health care clearinghouses, and health care providers) to protect the confidentiality of health care information that exists in electronic form. The Privacy Rule dramatically expands the class of IOs that are subject to federal legal requirements of confidentiality.

To some extent, the confidentiality promised by an IO is necessarily at risk. An IO cannot simply erect firewalls around its data, because the IO has a mandate to disseminate products based on these data. This mandate is based on an awareness that their data products contribute legitimate information to their clients. In a democratic and free market society, the client base of many IOs is broad. Statistical agencies, for example, not only provide data to guide government policy making, but they also provide data products to individuals, firms, non-governmental organizations, the media, and interest groups. As a most desirable result, public policy debate and decentralized economic decision making are informed. On the other hand, unintended consequences of dissemination can occur if the released information allows the confidentiality pledge to be compromised by a *data snooper*. The term *data snooper* refers to anyone with legitimate access to the data product and whose goals and methods in the use of the data are not consonant with the mission of the agency. Thus, a hacker who tries to break into a protected computer system is not a data snooper. Nor is a researcher who uses exploratory data analysis to discover statistical relationships. Other terms in the literature for “data snooper” include “data spy,” “intruder,” or “attacker.” Compromise of confidentiality by a data snooper constitutes a statistical confidentiality disclosure (Elliot and Dale 1999). Such a compromise occurs when the data dissemination permits a data snooper to gain illegitimate information about a respondent.

Ensuring confidentiality is not a simple task. For most of the census or survey data collected by statistical agencies, *deidentification*—removal of apparent identifiers like name, social security number, email address, etc. (although an obvious first step)—is not adequate to lower disclosure risk to an acceptable level (Paass 1988, Winkler 1998). Also, most health care information, such as hospital discharge data, cannot be anonymized through deidentification. The key reason that removing identifiers does not assure sufficient anonymity of respondents is that, today, a data snooper can get inexpensive access to databases with names attached to records. Marketing and credit information databases and voter registration lists are exemplars. Having this external information, the data snooper can employ sophisticated, but readily available, record linkage techniques. The resultant attempts to link an identified record from the public database to a deidentified record provided by the IO are often successful (Winkler 1998). With such a linkage, the record would be *reidentified*.

To publicly disseminate a data product safe from attack by a data snooper, an IO must go beyond deidentification; it must restrict the data by employing a disclosure limitation method. An easily interpreted and implemented method is to coarsen the data, essentially creating bins and counting the number of occurrences in the data. For example, recode income in increments of $5,000 and release a table giving, say, how many earned between $60,000 and $65,000. Coarsening provides a good example of an approach that while effective in lowering disclosure risk also lowers the data's utility. By fuzzing the target of a record linkage, such coarsening clearly makes reidentification through record linkage less likely. On the other hand, data utility becomes a problem with this coarsening to tabular form because releasing such tables no longer satisfies many users of statistical data. Coarsen gender, for instance, and you've lost the attribute entirely. Those data users who command the latest computer technology and who can make the most important research and policy contributions typically need data that are more distinguished. To be able to assess alternative disclosure limitation methods, we first need a framework for assessing how good a disclosure limitation procedure is.

In fulfilling their stewardship responsibilities, information organizations must manage the not easily resolved tension between ensuring confidentiality and providing access (Duncan, Jabine and de Wolf 1993, Kooiman, Nobel and Willenborg 1999, Marsh *et al* 1991). Resolving this tension requires policies under which an IO can disseminate data products that have both

1. high data utility U, so faithful in critical ways to the original data (analytically valid data), and
2. low disclosure risk R, so confidentiality is protected (safe data).

Statistical disclosure limitation techniques (Chowdhury *et al* 1999; Duncan and Lambert 1986; Zayatz *et al* 1999) provide classes of transformations that lower disclosure risk. Complicating the IO's task is the cornucopia of available statistical disclosure limitation methods, each with different impacts on data utility and disclosure risk. Major methods include suppressing attributes, swapping attributes, releasing only a sample of the population, topcoding, adding noise, various forms of aggregation, and cell suppression. General references to the literature in disclosure limitation include Duncan (2001), Duncan, Jabine and de Wolf (1993), Eurostat (1996), Fienberg (1994, 1997), Jabine (1993b), Mackie and Bradburn (2000), and Willenborg and de Waal (1996).

This article looks systematically at the simultaneous impact on disclosure risk and data utility of implementing disclosure limitation techniques and choosing their parameter values. A measure of statistical disclosure risk, R, is a numerical assessment of the risk of unintended disclosures following dissemination of the data. A measure of data utility, U, is a numerical assessment of the usefulness of the released data for legitimate purposes. Illustrative results using particular specifications for R and U are developed. In the next section, we introduce and exploit the *R-U confidentiality map*, the rudiments of which were presented by Duncan and Fienberg (1999) and further explored for categorical data by Duncan *et al.* (2001). By fully developing the concept of an R-U confidentiality map and applying it in some important contexts, this article provides a quantified link between R and U directly through the parameters of the disclosure limitation procedure. With an explicit representation of how the parameters of the disclosure limitation procedure affect R and U, the tradeoff between disclosure risk and data utility is apparent. With the R-U confidentiality map, information organizations have a workable new tool to frame decision making about data dissemination under disclosure limitation.

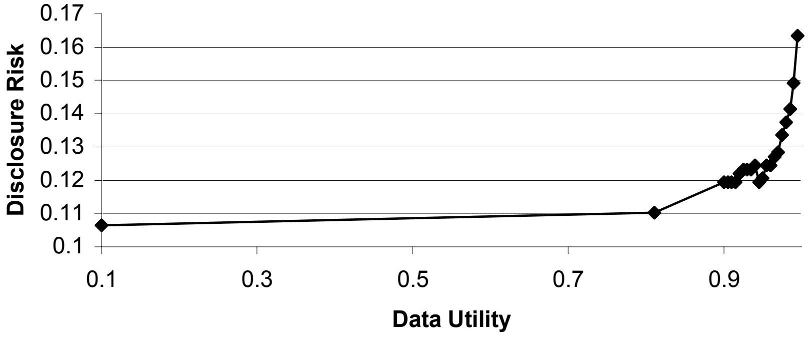
2. R-U Confidentiality Map

In its most basic form, an R-U confidentiality map is the set of paired values, (R, U), of disclosure risk and data utility that correspond to various strategies for data release. Typically, these strategies implement a disclosure limitation procedure, like masking through the addition of random error. Such procedures are determined by parameters, for instance, the magnitude of the error variance λ^2 for noise addition. As λ^2 is changed, a curve is mapped in the R-U plane. Visually, the R-U confidentiality map portrays the tradeoff between disclosure risk and data utility as λ^2 increases, and so more extensive masking is imposed.

Consider now an example of how a confidentiality problem can be recast in the form of the R-U confidentiality map. Take the disclosure risk R to be assessed as the percentage of records that can be correctly reidentified using record linkage software. As described in Moore

(1996), Winkler and Kim carried out a simulation experiment using Kim's (1986) method for adding noise to multivariate data. The records are perturbed so that the variance-covariance matrix is inflated by a factor of $1+\lambda^2$. They used the original database as a most conservative proxy for any external database that the data snooper may have. From the original database of 64,998 records of the Public Use File of the 1993 Annual Housing Survey, they extracted a target set of 771 records that had unusual categorical combinations. Then, using the procedure of Paass and Wasuchkuhn (1985), they sought to link them to particular ones of the 64,998 perturbed records. To develop an illustration of an R-U confidentiality map, we take the data user to estimate a mean with the data utility of the perturbed data as $1/\ (1+\lambda^2)$, which is proportional to the reciprocal of the variance of the sample mean of the perturbed data. We define the disclosure risk to be the expected proportion of the 771 records that can be correctly reidentified. This quantity was estimated in the Kim and Winkler experiment. The resulting R-U confidentiality map is as Figure 1.

Figure 1. R-U Confidentiality Map for Kim's Method with Housing Data



Note the following for Figure 1. First, the map is not completely smooth because of empirical variation in the number of correct reidentifications. Second, working from the right of the map, the points correspond to the disclosure-limitation parameter λ^2 increasing from 0.0049 to 0.2333, and then jumping for the last point to 9. Thus, the disclosure risk drops only slightly for very substantial increases in the added noise parameter λ^2, while data utility plummets. This suggests that there is little value in having λ^2 increase beyond about 0.2 for these data. Third, this R-U confidentiality map illustrates a conservative case for two reasons: (1) the method of Paass and Wauschkuhn (1985) is near a worst case scenario for what external database a data snooper can use for record linkage, and (2) the target values were chosen to have unusual attribute values, which presumably makes them easier to reidentify.

Data utility is a familiar concept to statisticians and empirical researchers. It has been recognized as a consideration in disclosure limitation by a few others (e.g., Kim 1986, Marsh *et al* 1991, and Skinner 1990). Kamlet, Klepper and Frank (1985) give examples that show how disclosure limitation techniques can seriously alter relationships. Agarwal and Srikant (2000) develop a metric for data utility based on the width of the 95% prediction interval. We explore some additional possibilities here. On the other hand, disclosure risk is less familiar and so we next investigate this concept in some generality.

Present practice by IOs in assessing tradeoffs between disclosure risk and data utility is primarily heuristic. Recommendation 6.2 of the National Academy of Sciences Panel on Confidentiality and Data Access (Duncan, Jabine and de Wolf 1993) advises the development of theoretical foundations for such determinations. Approaches to disclosure risk based on a decision-theoretic characterization of the data snooper are developed by Duncan and Lambert (1986, 1989), Lambert (1993), Little (1993), Mokken *et al* (1992) and, more fully, for both the data snooper and the IO by Trottini (2001). The idea is to view the actors as decision makers, who take actions in light of their perceptions of probabilities and consequences:

1. The data snooper can choose (or not) to make identifications and draw inferences on the basis of the released data product, and

2. The IO chooses a statistical disclosure limitation method to deter the data snooper.

From this perspective, disclosure risk depends on the decision structure—probabilities and utilities of consequences—of the data snooper and IO. There are, however, two important complications from the usual decision model: (1) the IO has only its own perceptions of the decision structure of the data snooper, and (2) the IO must cope with multiple data snoopers. The solution to (2) is general—protect against the worst. The solution to (1) is also general—put yourself in the shoes of the data snooper. Implementing both of these solutions reduces the problem back to the decision structure of a single individual.

Disclosure risk is viewed from the perspective of the IO. Most generally, it has three components:

1. The ease with which the data snooper can make inferences about a target value τ based on the released data product,
2. The consequences to the data snooper of own inferences—none, correct or incorrect, and
3. The consequences to the IO of inferences by the data snooper.

Once the form of the data utility U and the disclosure risk R have been appropriately specified, our task is to determine how R and U are related to the parameter values of the disclosure limitation methods under consideration. This gives us the R-U confidentiality map. We next do this for additive noise.

3. R-U Confidentiality Map For Additive Noise

In this section, we demonstrate how an R-U confidentiality map can be constructed to examine the impact of a disclosure limitation procedure that has received substantial attention—additive noise (e.g., Kim and Winkler 1995; Duncan and Mukherjee 2000). For purpose of illustration, we take the data to be a random sample X_1, …, Xn from a univariate population with mean μ and standard deviationσ, the data utility U to be the reciprocal of the data user's mean squared error in estimating the population mean μ, 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. In practice, the form of R and U should be tailored to the particular situation at hand. In developing measures of the disclosure risk R, we consider three different knowledge states, depending on the group to which the data snooper can isolate the target:

1. *Population.* The target τ has the same distribution as the X_i;
2. *Sample.* The target τ is one of the X_i's, (i.e., is in the sample); or
3. *Record.* The target τ is not only in the sample, but the data snooper has enough external information to be able to identify (link to) the specific masked record corresponding to X_i.

The first case is most 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 where the data snooper can isolate the target to the sample is most appropriate when the data are a census or a near census. The third case is most appropriate when the data snooper has extensive external information that would permit record linkage to identify record i as the target. Results will be developed and discussed for all three states of knowledge of the snooper 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 not atypical 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

For the realized values, x_1, …, x_n, the masked data has the additive noise form,

$Y_i = x_i + \varepsilon_{i,} \varepsilon_i \sim iid(0, \lambda^2),\ i = 1, \ldots, n.$.

Data Utility: The data user estimates the population mean μ using

$\hat{\mu} = \bar{Y}$, the sample mean of the masked data. Therefore, $E(\hat{\mu}) = \mu,\ Var(\hat{\mu}) = \frac{1}{n}\left(\sigma^2 + \lambda^2\right)$, and the data utility is $U = \frac{n}{\sigma^2 + \lambda^2}$.

Disclosure Risk: The first two states of knowledge of the data snooper, assuming the snooper's goal is to compromise a specific entity, have the same disclosure risk. In both states the data snooper is simply after a specific target value τ and will use $\hat{\tau} = \bar{Y}$ as the estimator for τ. This gives risk of

$$R = \frac{1}{E(\tau - \hat{\tau})^2} = \frac{n}{\sigma^2 + \lambda^2 + n(\mu - \tau)^2}. \tag{1}$$

Given the risk specified by Equation (1), the IO can determine what entities lead to the maximum risk across either the sample or the entire population. Since R is maximized for $\tau = \mu$, the IO can decide whether disclosure of attribute values near the mean pose serious consequences. If, instead, the IO felt that less typical values were of more concern, the IO can consider targets at the population average value for $(\mu - \tau)^2$, which is σ^2. This gives a disclosure risk of

$$R = \frac{n}{(n+1)\sigma^2 + \lambda^2}. \tag{2}$$

With similar motivation, the IO can consider a sample average value for $(\mu - \tau)^2$. This gives a disclosure risk of

$$R = \frac{n}{\sigma^2 + \lambda^2 + \sum_{i=1}^{n}(\mu - x_i)^2}. \tag{3}$$

The third state of knowledge for the data snooper is the worst case with respect to maximum risk. In this state, the snooper is able to identify the masked record that corresponds exactly to the target τ. Here, if the snooper uses $\hat{\tau} = Y_i = \tau + \varepsilon_i$, the risk would be

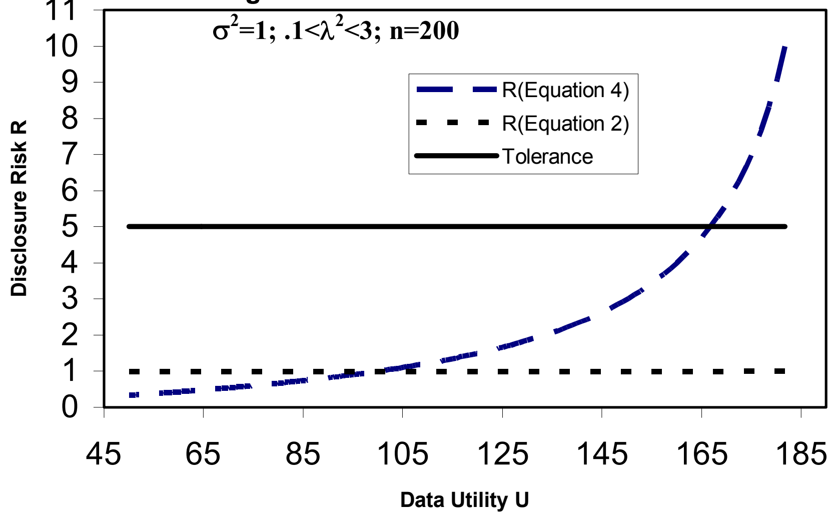
$$R = \frac{1}{E(\tau - \hat{\tau})^2} = \frac{1}{E(\varepsilon)^2} = \frac{1}{\lambda^2}. \tag{4}$$

Whatever the knowledge state of the data snooper, the disclosure risk without data masking is found by setting $\lambda^2 = 0$. Note that R is infinite if the data snooper can link τ with certainty to a particular record i and no noise has been introduced. Thus, in circumstances where record linkage is feasible, release of the original data would have substantial disclosure risk and pose too much of a threat to confidentiality. Before release, the data would have to be masked.

Is the data snooper always better off, when knowing the target's index i, using $\hat{\tau} = Y_i$ to assess the target value $\tau = x_i$? Comparing Equations (2) and (4), we see that the data snooper actually gains by using $\overline{Y}$ whenever the value of the error variance $\lambda^2 > \left(\frac{n-1}{n+1}\right)\sigma^2$. Thus, by adding sufficient noise—of the magnitude of the population variance σ^2, the IO can reduce the advantage the data snooper has through record linkage.

Displayed in Figure 2 is an R-U confidentiality map for two risk measures in this example. The figure displays the impact on data utility and disclosure risk for changes in the disclosure limitation parameter λ^2, under both the assumption that the data snooper knows the index of the target (Equation (4)) and the assumption that the data snooper does not (Equation (2)). For illustration, when the data snooper knows the target's index and the maximum tolerable disclosure risk is 5, the optimal value for the variance of the noise addition is $\lambda^2 = 0.21$, yielding a data utility of 165. If the data snooper does not know the target's index, the disclosure risk is low enough that no noise addition is necessary. This comparison under the two knowledge states shows how valuable policies are that make it difficult for a data snooper to know the index of a target. For example, if the data are a sample from some population, even an exact match on record linkage with some external database with identifiers is insufficient to guarantee that the linked record corresponds to the target individual. Thus sampling gives some confidentiality protection.

Figure 2: R-U Confidentiality Map for Additive Noise: Knowledge States 1 and 2 vs. 3



3.2. Target at a Percentile Point

Let's think like a data snooper and recognize the fact that extreme or outlying data values may present better—that is, more easily compromised—targets. For the agency, this vulnerability is doubly serious because targets with atypical values often pose more serious consequences. For such targets, the snooper can use estimators of a specified percentile point or estimators of an extreme value for 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. τ is the p^{th} *population* percentile point; or
2. Record knowledge. τ is known to be in the sample and to be the p^{th} *sample* percentile point.

We establish notation and structure: Let $X_1, X_2, \ldots, X_n \sim iid\left(\mu, \sigma^2\right)$, with the data being the realizations, $x_1, \ldots, x_n$. Denote the p^{th} population percentile point as $\xi_{Xp} = \mu + l_{Xp}\sigma,$ so that $F_X(\xi_{Xp}) = P(X \le \mu + l_{Xp}\sigma) = p.$ Apply independent noise addition, as before, with the released values $Y_i = x_i + \varepsilon_i,\ \varepsilon_i \sim iid(0, \lambda^2),\ i = 1, \ldots, n.$ Finally, denote the p^{th} population percentile point of the masked data as $\xi_{Yp} = \mu + l_{Yp}\sqrt{\sigma^2 + \lambda^2}$, so that

$F_Y(\xi_{Yp}) = P(Y \le \mu + l_{Yp}\sqrt{\sigma^2 + \lambda^2}) = p.$

We assume the data snooper uses an obvious strategy of $\hat{\tau} = \hat{\xi}_{Yp} = Y_{(np)}$ as the attack estimator for either the population or sample percentile, depending on the state of knowledge above. This estimator is biased when the data has been altered by additive noise, but without additional knowledge of the details of the model, the snooper cannot reduce the bias of his estimate (Stefanski and Bay 1996). Consider a value of p so that np is an integer. With a strict monotonicity assumption on the distribution function $F_X(x)$, the percentile estimator has the asymptotic distribution (see Mood, Graybill, and Boes, 1963, p. 257),

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

If the data snooper's state of knowledge is that τ is the p$^{\text{th}}$ *population* percentile, then the disclosure risk is

$$R = \frac{1}{E(\tau - \hat{\tau})^2} \approx \frac{1}{\frac{p(1-p)}{n[f_Y(\xi_{yp})]^2} + \left(l_{Yp}\sqrt{\sigma^2 + \lambda^2} - l_{Xp}\sigma\right)^2}. \tag{5}$$

From this formula, we can see circumstances under which the IO can adequately lower its disclosure risk by setting λ^2 sufficiently large. The dual of R low is that the mean squared error for the data snooper is high. 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, $\hat{\xi}_{Yp}$. 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 snooper could make an adjustment for the bias in $\hat{\xi}_{Yp}$. The data snooper has two possible sources of information about the value of λ: (1) the IO could have revealed the value it used in masking, or (2) based on experience with similar data, the 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 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 (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 agency's use of additive noise has made inoperative direct use of the knowledge that the target is at a particular percentile point. Comparing Equations (2) and (5) gives circumstances when the data snooper should switch to $\bar{Y}$.

Consider the second state of knowledge where the snooper knows that the target is the *sample* pth percentile point, $\tau = x_{(np)}$. Still using $\hat{\tau} = \hat{\xi}_{Yp} = Y_{(np)}$, the disclosure risk is

$$R = \frac{1}{E(\tau - \hat{\tau})^2} = \frac{1}{Var(\hat{\tau}) + (\tau - \xi_{Y_p})^2}. \tag{6}$$

Equation (6) yields no insight into how disclosure risk behaves over the range of p. Restating Equation (6) as

$$R = \frac{1}{\sum_{i=1}^{n} E_{\varepsilon}[P(\hat{\tau} = x_i + \varepsilon_i)(\tau - x_i - \varepsilon_i)^2]}. \tag{7}$$

we see that the disclosure risk R is maximized when the npth masked value corresponds exactly to the npth original data point, or $\hat{\tau} = x_{(np)} + \varepsilon_{(np)}$. In this case, R=1/$\lambda^2$. This is equivalent to the case where the data snooper is able to link exactly to the target's masked value (see Equation (4)), which 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.

To demonstrate the use of the R-U confidentiality map framework for percentile estimation attacks (Equation (5)) versus mean attacks (Equation (2)) and direct record linkage (Equation (4)), we need to make some additional distributional assumptions. We assume X_i and ε_t are normally distributed. Then, $l_{xp}=l_{yp}=z_p$, and $f_x(\xi_{xp}) = \frac{\varphi(z_p)}{\sigma}$ and $f_y(\xi_{yp}) = \frac{\varphi(z_p)}{\sqrt{\sigma^2 + \lambda^2}}$, where z_p is the pth percentile of a standard normal distribution and $\varphi(\cdot)$ is the standard normal density function. Equation (5) becomes

$$R \approx \frac{1}{\left(\dfrac{p(1-p)}{n[\varphi(z_p)]^2}\right)(\sigma^2+\lambda^2)+z_p^2\left(\sqrt{\sigma^2+\lambda^2}-\sigma\right)^2}. \qquad (8)$$

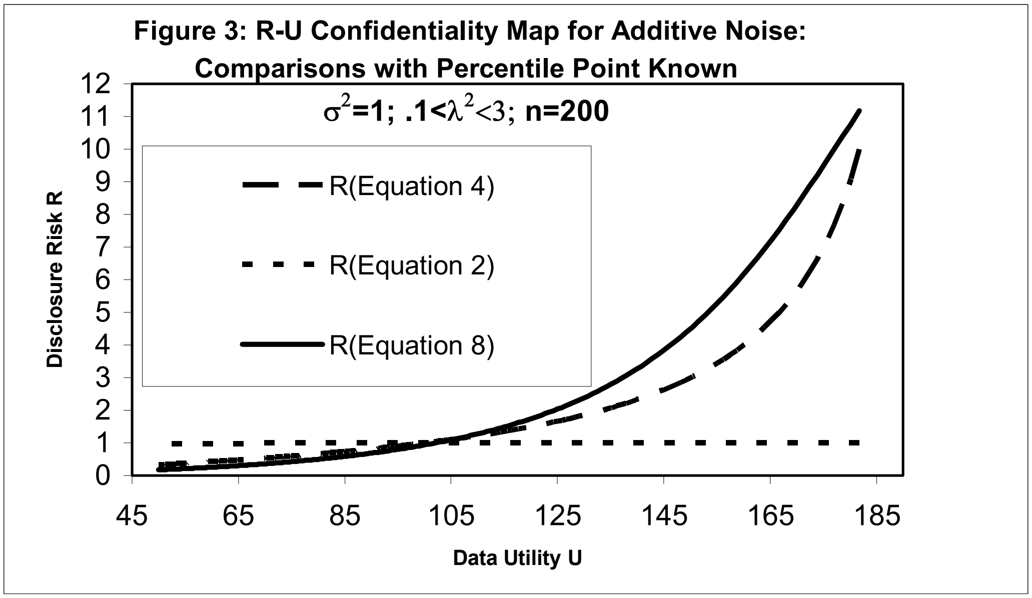


Figure 3: R-U Confidentiality Map for Additive Noise: Comparisons with Percentile Point Known

Figure 3 displays the impact on data utility and disclosure risk for changes in the disclosure limitation parameter λ^2, under the assumption that the data snooper knows that the 99th population percentile is the target (Equation (8)), only knows the target is from the same population as the sample (Equation (2)), and knows how to link the masked value to the target (Equation (4)). Note that the R-U confidentiality maps under the different knowledge states for the data snooper cross. Knowing that the target is at the 99th percentile is of little value to the data snooper if substantial noise is added to the data, but is of great value if little noise is added to the data.

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, let $X_1,...,X_n \sim N(\mu,\sigma^2)$ with realizations x_1, …, x_n. Mask through $Y_i = x_i + \varepsilon_i$, $\varepsilon_i \sim N(0,\lambda^2)$, independently. If the 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 $E(\tau-\hat{\tau})^2 = E(x_{(n)} - Y_{(n)})^2 = Var(Y_{(n)}) + (x_{(n)} - \mu_{Y_{(n)}})^2$, where $\mu_{y(n)}$ is the mean of the maximum order statistic $Y_{(n)}$. For any finite *n*, this expression can be evaluated using a table of moments of normal order statistics. For large samples, we can appeal to the classic results of Fisher and Tippett (1928) to show that in large samples, $\mu_{y(n)} \approx \mu + \sqrt{\sigma^2+\lambda^2}K_{n1}$, where $K_{n1} = \sqrt{2\log n} - \dfrac{\log\log n + \log 4\pi - 2\gamma}{2\sqrt{2\log n}}$ and $\gamma = 0.57722$ is Euler's constant; and $Var(Y_{(n)}) \approx (\sigma^2+\lambda^2)K_{n2}$, where $K_{n2} = \dfrac{\pi^2}{12\log n}$. [An accessible reference for these results is Cramér (1946), p. 376.] Thus,

$$R = \frac{1}{E(x_{(n)} - Y_{(n)})^2} \approx \frac{1}{K_{n2}(\sigma^2+\lambda^2) + (x_{(n)} - (\mu + K_{n1}\sqrt{\sigma^2+\lambda^2}))^2}. \tag{9}$$

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

To demonstrate the R-U confidentiality map for this example we will substitute $x_{(n)} \approx \mu + \sigma K_{n1}$ into Equation (9). This gives a risk of

$$R \approx \frac{1}{K_{n2}(\sigma^2+\lambda^2) + K_{n1}^2(\sqrt{\sigma^2+\lambda^2} - \sigma)^2} \tag{10}$$

Note the similarity between Equation (10) and Equation (8); both numerators are linear combinations of $\sigma^2+\lambda^2$, which is the variance of a masked observation, and $(\sqrt{\sigma^2+\lambda^2}-\sigma)^2$, 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 (10)—so the target is an extreme—goes to zero, while the disclosure risk given by Equation (8)—so the target is a percentile point—goes to a positive value.

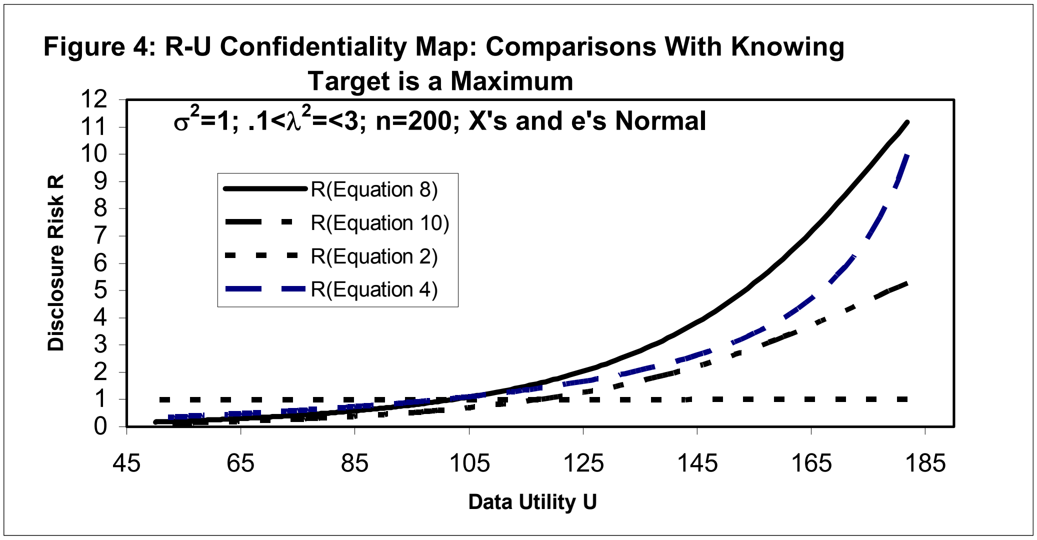


Figure 4: R-U Confidentiality Map: Comparisons With Knowing Target is a Maximum

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 (10)), knows the 99th population percentile is the target (Equation (8)), only knows the target is from the same population as the sample (Equation (2)), and knows how to link the masked value to the target (Equation (4)). 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

The previous sections developed the general theory of the R-U confidentiality map and showed how it provides qualitative insights for classes of disclosure limitation procedures. In this section we show how an organization can produce and make use of an R-U confidentiality map for a particular database. Using a real-life example of both practical size and realistic complexity, we detail how this *empirical R-U confidentiality map* can be used to:

- inform the IO about whether or not proposed disclosure limitation methods are adequate in lowering disclosure risk and maintaining data utility,
- facilitate comparisons between various disclosure limitation methods, and
- examine the risk of particular types of data snooper knowledge.

Analytical methods, such as those developed in the previous sections, can be used to investigate general properties of disclosure limitation methods and snooper strategies. In practice, nonetheless, the probabilistic structure underlying the analytic development may not be fully adequate to depict consequential features of the actual data. Also, it may be difficult to derive the impact of disclosure limitation methods on disclosure risk and data utility. The empirical approach laid out in this section can, therefore, be helpful to an agency considering how to disseminate data products from a specific database.

To illustrate the empirical approach, we use data from a survey by the NCES—the Teacher Followup Survey (TFS) of 1994-95. The attribute we examine is total household income (identified as TFS376 in the survey documentation). The data were obtained from a sample of teachers, first interviewed in 1993-94 under the School and Staffing Survey (SASS). The SASS and TFS were part of a series of surveys conducted by the U.S. Bureau of the Census for NCES to provide a description of the nation’s public and private schools. The goal of the TFS was to investigate attrition rates for teachers, and to elicit characteristics and attitudes of leavers and non-leavers from the profession.

In our illustration, the IO needs to assess the utility of the data for estimation of the mean household income μ_x of the population of part-time private school teachers. Summary statistics of the data for household income are shown in Table 1 (the value of the sample size n is considered confidential by NCES).

Table 1. Summary Statistics for the 1993-94 Teacher Followup Survey (TFS376): Household Income of Part-Time Private School Teachers

n	Sample Mean	Sample SD	Min	1st %-ile	10th %-ile	90th %-ile	99th %-ile	Max
NA	\$20.1K	\$13.3K	\$2K	\$2K	\$7.2K	\$35K	\$70K	\$95.2K

The goal of the data snooper is to infer the household income of a particular target record. The IO is studying normal noise addition as a disclosure limitation method. However, since income is bounded below by 0, the IO truncates the value of the masked variable at 0. Thus, the actual household income for the ith record, x_i, is reported in the released file as

$$Y_i = \max(0, x_i + \varepsilon_i), \tag{11}$$

where $\varepsilon_\iota \sim N(0, \lambda^2)$ for each i. Because the masking in Section 3 did not include truncation, the expressions for data utility and disclosure risk presented there are not directly applicable. Nevertheless, we can show that the agency can construct an R-U map for this method—or, indeed, any proposed method of disclosure limitation—by obtaining R and U empirically.

To obtain the empirical R-U confidentiality map in the present example, the agency can use the form of model specified by Equation (11) to simulate the masking process. For each of a range of values of λ^2, the agency can generate a number, say M, of masked datasets. From these simulated datasets, the agency can estimate the disclosure risk $R = 1/E(\tau - \hat{\tau})^2$ and the data utility $U = E(\overline{Y} - \mu_x)^2$.

We conducted such a simulation with $M = 200$ and several values of λ^2 ranging from 5 percent to 100 percent of the sample variance, so $0.05\,S_x^2$ to $1.0\,S_x^2$. We examined several different types of targets and two different strategies based on assumptions about the knowledge of the data snooper. The target types were $\tau =$ maximum, minimum, and p^{th} sample percentile for $p = 0.01, 0.10, 0.90$, and 0.99. For the data snooper, the strategies based on two record knowledge states were:

1. *Index knowledge:* $\hat{\tau} = Y_t$, where t is the index of the record in the unmasked sample which corresponds to the largest, smallest, or p^{th} sample percentile, and
2. *Position knowledge:* $\hat{\tau} = \max_{1 \le i \le 177}(Y_i)$, $\min_{1 \le i \le 177}(Y_i)$, or $Y_{(np)}$.

Under the first strategy, the data snooper uses knowledge of the index of the target; it is the same as the case where the data snooper has enough external information to be able to identify (link to) a specific masked record. Under the second strategy, the data snooper uses own knowledge of the position of the target in the unmasked data to assess an attribute value from the masked data.

Our estimate of the snooper's MSE $E(\tau - \hat{\tau})^2$ is, for each value of λ^2,

$$\frac{1}{200}\sum_{m=1}^{200}(\tau - \hat{\tau}_m)^2\,, \tag{12}$$

where $\hat{\tau}_m$ is the snooper's prediction of the target in replicate m (m=1, ..., M=200). In Section 3.1, the data user's utility was taken to be $U = 1/E(\overline{Y} - \mu_x)^2 = n/(\sigma^2 + \lambda^2)$. In this case, the truncation of the variable at the natural bound of 0 introduces a bias and reduces the variance so that this expression for the user's MSE, $E(\overline{Y} - \mu_x)^2$, is no longer accurate. Instead, we have

$$E(\overline{Y} - \mu_x)^2 = Var(\overline{Y}) + (\mu_y - \mu_x)^2\,, \tag{13}$$

which can be estimated for each value of λ^2 by

$$\frac{1}{200}\sum_{m=1}^{200}\frac{s_m^2}{n}+(\overline{\overline{Y}}-\overline{x})^2 \ , \tag{14}$$

where s_m^2 is the sample variance for the mth masked replicate, $\overline{\overline{Y}}$ is the average of the 200 sample means $\overline{Y}_m$, and $\overline{x}$ is the mean of the x_i values in the original sample. The empirically derived values of R and U are then taken as the reciprocals of the expressions in (12) and (14).

Figure 5. Empirical R-U Confidentiality Map: Comparing Risk When the Data Snooper Knows the Targeted Record to be a Maximum or a Minimum
1994-95 TFS, Private Schools, Part-Time Teachers

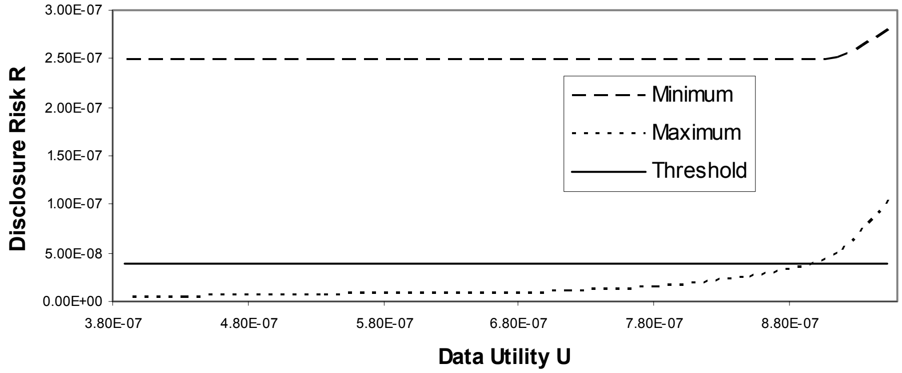
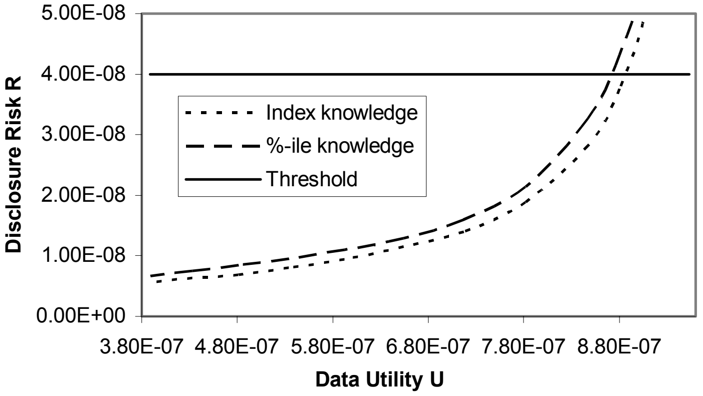


Figure 5 shows the empirical R-U confidentiality map for the case in which the data snooper knows the target to be the maximum or the minimum in the sample. The risk threshold has been set at 4.0E-08, which can be regarded as an average tolerable error of no more than \$5K (since $1/5000^2$ = 4.0E-08). If the snooper uses $\hat{\tau} = \min_{1\le i\le n}(Y_i)$, then there is no value of λ^2 that will produce a risk this low. This is because truncation of the masked income value at 0 limits the magnitude of the downside error to \$2K, which has an associated risk value well below the allowable \$5K. By contrast, if the target were known to be the maximum, then a choice of $\lambda^2 = 0.12\,S_y^2$ is sufficient to meet the threshold. This would lead to a data utility for the analyst of about 8.9E-07, or an efficiency of 8.9/10.0 = 89%, when compared to the unmasked.

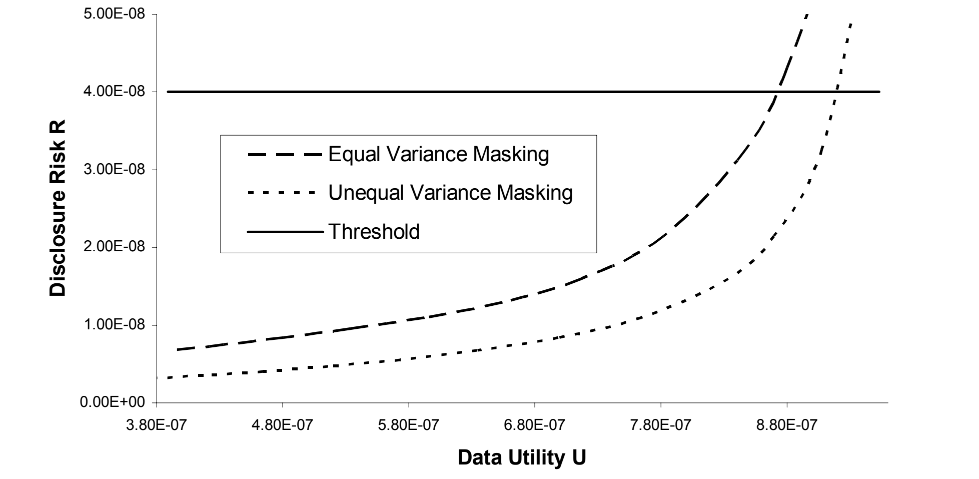
Figure 6 is an empirical R-U confidentiality map that compares the disclosure risk for changes in the disclosure limitation parameter λ^2, under different kinds of knowledge the snooper might have about the same record. In particular, it shows the risk for the cases when the data snooper knows the target is the 99th sample percentile ($70K in for this sample) and when the snooper can link the masked value to the target. This map illustrates that the disclosure risk from knowledge of the percentile is greater than that from knowledge of the index of the record. This ordering was true for all of the targets in this example. Thus, an adequate masking strategy must depend on what knowledge the snooper is likely to have about the target record.

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)
1994-95 TFS, Private Schools, Part-Time Teachers



Finally, Figure 7 shows how an empirical R-U confidentiality map might be used to help the agency compare two alternative masking strategies. “Equal variance masking” is the one implemented as shown in Equation (11). In “unequal variance masking”, the variance of the noise is doubled for sensitive data, which in this case is defined as income values exceeding $35K. Here we see that if the goal were to protect against this percentile knowledge and maintain the maximum disclosure risk specified, we could do so with unequal variance masking and keep higher data utility. Of course, in practice an agency must be prepared to protect against a variety of snooper targets and strategies, so the decision about the type of masking to use could not be made based on a single R-U confidentiality map. Nonetheless, these maps do provide a tool for addressing the problem.

Figure 7. Empirical R-U Confidentiality Map: Comparing Risk for Two Masking Methods when Target is Known to be at 99th Percentile 1994-95 TFS, Private Schools, Part-Time Teachers



5. Conclusions

This work addresses the need that information organizations have to disseminate useful data while keeping low the risk of statistical confidentiality disclosure. Recognizing that deidentification of data is inadequate to protect confidentiality against attack by a data snooper, agencies restrict the data they release to general data users. Typically, these restricted data procedures have involved transformation or masking of the original, collected data through such devices as adding noise, topcoding, data swapping, and recoding. This paper gives a framework for the determination of the parameter values of a disclosure limitation procedure and for the comparison of disclosure limitation procedures. The framework focuses on the tradeoff between data utility and disclosure risk. Examples are provided which illustrate the concepts by putting forth quantitative measures of both data utility and disclosure risk. These measures permit the agency to consider the tradeoffs between providing more useful data to users and lowering the risk to confidentiality. For work with a particular database, we have shown how to use simulation methods to develop an empirical R-U confidentiality map. Applications were made to survey data from the National Center for Education Statistics.

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 while George Duncan was on leave from Carnegie Mellon University at Los Alamos National Laboratory. 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."

References

Agarwal, R. and Srikant, R. (2000) Privacy-preserving data mining. *Proceedings of the 2000 ACM SIGMOD on Management of data,* May 15-18, Dallas, Texas.

Chowdhury, S. D., Duncan, G. T., Krishnan, R., Roehrig, S. F., and Mukherjee, S. (1999) Disclosure detection in multivariate categorical databases: Auditing confidentiality protection through two new matrix operators. *Management Science* **45** 1710-1723.

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

Duncan, G. T. (2001) Confidentiality and statistical disclosure limitation. *International Encyclopedia of the Social and Behavioral Sciences.* To appear.

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* (J. Lane, editor)

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. (1986) Disclosure-limited data dissemination (with discussion). *Journal of the American Statistical Association*. **81** 10-28.

Duncan, G. T. and Lambert, D. (1989) The risk of disclosure of 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.

Eurostat (1996) Manual on Disclosure Control Methods. Luxembourg: Office for Publications of the European Communities.

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.

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

Kamlet, M. S., Klepper, S. and Frank, R. G. (1985) Mixing micro and macro data: Statistical issues and implication for data collection and reporting. *Proceedings of the 1983 Public Health Conference on Records and Statistics*, U.S. Department of Health and Human Services, Hyattsville, MD.

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. (1995) Masking microdata files. Proceedings of the Section on Survey Research Methods, American Statistical Association. (???)

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.

Little, R. J. A. (1993) Statistical analysis of masked data. *Journal of Official Statistics* **9** 407-426.

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.

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.

Mokken, R. J., Kooiman, P., Pannekoek, J, and Willenborg, L. C. R. J. (1992) Disclosure risks for microdata. *Statistica Neerlandica.* **46** 49-67.

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

Moore, R. A. (1996) Controlled data-swapping techniques for masking public use microdata sets. *Statistical Research Division Report Series, RR 96-04.* Washington, DC: U.S. Bureau of the Census.

Paass, G. and Wasuchkuhn, U. (1985) Datenzugang, datenschutz, und anonymisierung. *Analyysepotential 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.

Skinner, C. J. (1990) Statistical disclosure issues for census microdata. Paper presented at International Symposium on Statistical Disclosure Avoidance, Voorburg, The Netherlands, December 13.

Stefanski, L. and Bay, J.M. (1996) Simulation extrapolation deconvolution of finite population distribution function estimators, *Biometrika* 83, 407-417.

Trottini, Mario (2001) A decision-theoretic approach to data disclosure problems. Joint ECE/Eurostat Work Session on Statistical Data Confidentiality. Skopje, The Former Yugoslav Republic of Macedonia, March 14-16.

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

Winkler, W. E. (1998) Re-identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata. *Research in Official Statistics*, **1**, 87-104.

Zayatz, L. V., Massell, P. and Steel, P. (1999) Disclosure limitation practices and research at the U.S. Census Bureau. *Special issue on statistical disclosure control. Netherlands Official Statistics* **14** 26-29.