A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality

A. F. KARR, C. N. KOHNEN, A. OGANIAN, J. P. REITER, and A. P. SANIL

When releasing data to the public, statistical agencies and survey organizations typically alter data values in order to protect the confidentiality of survey respondents' identities and attribute values. To select among the wide variety of data alteration methods, agencies require tools for evaluating the utility of proposed data releases. Such utility measures can be combined with disclosure risk measures to gauge risk-utility tradeoffs of competing methods. This article presents utility measures focused on differences in inferences obtained from the altered data and corresponding inferences obtained from the original data. Using both genuine and simulated data, we show how the measures can be used in a decision-theoretic formulation for evaluating disclosure limitation procedures.

KEY WORDS: Confidentiality; Disclosure; Disclosure risk; Microdata; Record linkage; Statistical disclosure limitation; Utility.

1. INTRODUCTION

A central mission of many statistical agencies and survey organizations is to disseminate microdata, that is, individual data records, to researchers or the public. Dissemination of microdata greatly benefits society, as well as facilitates research and advances in economics, public health, sociology, and many other areas of knowledge. Disseminating microdata—as compared, for example, to remote access servers (Gomatam, Karr, Reiter, and Sanil 2005a)—benefits researchers, who may perform a wide variety of analyses.

Usually, however, data disseminators cannot release microdata as collected, because doing so would reveal respondents' identities or values of sensitive attributes. Agencies that fail to protect confidentiality may be in violation of laws such as the recently enacted Confidential Information Protection and Statistical Efficiency Act of 2002 (Wallman and Harris-Kojetin 2004) in the U.S. Additionally, if confidentiality is compromised, organizations may lose the trust of the public, so that potential respondents are less willing to give accurate answers, or even to participate in surveys.

To reduce disclosure risks, data disseminators typically remove key identifiers and/or alter values of sensitive attributes before releasing data. For example, they recode variables, releasing ages or incomes in aggregated categories. Instead, they may swap data values for selected records, for example, switching the sexes of some men and women in the data, in hopes of discouraging users from matching, since matches may be based on incorrect data. Or, they add noise to numerical data values to reduce the likelihood of exact matching on key variables or to distort the values of sensitive variables. Indeed, virtually all public use data releases have undergone some form of statistical disclosure limitation (SDL).

SDL methods can be implemented with differing degrees of intensity. Generally, increasing the amount of alteration decreases the risk of disclosure, but it also decreases the accuracy of inferences obtainable from the released data, often referred to as data utility (Willenborg and de Waal 2001).

Although there is a plethora of SDL techniques, there exist few principled methods for selecting which technique, and with what degree of intensity, to employ in a particular setting. Formally or informally, most selection methods are based on trading off some notion of disclosure risk for some notion of data utility, often referred to as data quality (Karr, Sanil, and Banks 2006). Such formulations have been described for data swapping (Gomatam, Karr, and Sanil 2005b), regressions (Gomatam et al. 2005a), tabular data (Dobra, Fienberg, Karr, and Sanil 2002; Dobra, Karr, and Sanil 2003; Duncan and Fienberg 1999; Duncan et al. 2001) and other settings (Domingo-Ferrer, Mateo-Sanz, and Torra 2001; Duncan, Keller-McNulty, and Stokes 2004).

In a formal risk-utility formulation, each candidate release R—which is a function of the original database $\mathcal{D}_{\text{orig}}$ and possibly exogenous randomness—is characterized by a quantified *disclosure risk* $\text{DR}(R)$ [which may be that of either identity or attribute disclosure (Duncan and Lambert 1989)] and *data utility* $\text{DU}(R)$. The actual release $\mathcal{D}_{\text{rel}}$ can be selected from the candidates in one of two ways. The first is to maximize utility subject to an upper bound on risk, by solving an optimization problem of the form

$$\mathcal{D}_{\text{rel}} = \arg\max_{R \in \mathcal{R}} \text{DU}(R), \tag{1}$$

where $\text{DR}(R) \leq \alpha$, and where $\mathcal{R}$ is the set of all candidate releases.

The second, and more flexible, approach is to define *risk-utility frontiers* using the partial order $\preceq_{\text{RU}}$ defined by

$$R_1 \preceq_{\text{RU}} R_2 \Leftrightarrow \text{DR}(R_2) \leq \text{DR}(R_1)$$

and

$$\text{DU}(R_2) \geq \text{DU}(R_1). \tag{2}$$

A. F. Karr is Director, National Institute of Statistical Sciences, P.O. Box 14006, Research Triangle Park, NC 27709 (E-mail: *karr@niss.org*). C. N. Kohnen is Assistant Professor, Duke University, Durham, NC. Currently at St. Olaf College, Northfield, MN 55057. A. Oganian is Postdoctoral Fellow, National Institute of Statistical Sciences, Research Triangle Park, NC 27709. J. P. Reiter is Assistant Professor, Duke University, Durham, NC 27708. A. P. Sanil is Senior Research Biostatistician, Global Biometric Sciences, Bristol-Myers Squibb Company, Princeton, NJ 08543. This research was supported by NSF grant IIS–0131884 to the National Institute of Statistical Sciences. Any opinions, findings, and conclusions or recommendations expressed in this article are those of the authors and do not necessarily reflect the views of the National Science Foundation.

When $R_1 \preceq_{\text{RU}} R_2$, the R_2 is preferred to R_1 because it has both lower disclosure risk and higher utility. Only candidate releases on the risk-utility frontier of maximal elements of $\mathcal{R}$ with respect to the partial order (2) need be considered further: for any other candidate, some element of the frontier has lower risk *and* higher utility. Calculation of the frontier can be done using existing algorithms for finding the maxima in a set of vectors (Kung, Luccio, and Preparata 1975).

While there has been much work on developing measures of disclosure risk (e.g., Duncan and Lambert 1986, 1989; Lambert 1993; Fienberg, Makov, and Sanil 1997; Skinner and Elliot 2002; Reiter 2005a), there has been comparatively little work on developing measures of data utility, and so this article outlines a framework for defining and comparing measures of data utility. Section 2 outlines the problem and defines utility measures that range from the very specific but very narrow—focused on one analysis of the data—to the very broad, but correspondingly blunt. Section 3.2 presents, in effect, a case study in using utility measures to select SDL methods. Because a particular database may not yield generalizable insights, Section 3.3, using simulated data, shows how the utility measures can be used to evaluate the characteristics of SDL methods across differing data structures. A concluding discussion is in Section 4.

2. UTILITY MEASURES

We begin with a general discussion of utility measures (Sec. 2.1), and then we introduce the three measures studied in this article (Secs. 2.3 and 2.2).

2.1 Generalities

Data utility measures should be linked to the types of analyses done on the released data, and that at some level they must measure the fidelity of analyses performed on the released database $\mathcal{D}_{\text{rel}}$ to the same analyses performed on the original database $\mathcal{D}_{\text{orig}}$. In a purely abstract sense, these measures are of the form $d(\mathcal{D}_{\text{rel}}, \mathcal{D}_{\text{orig}})$, where d is some possibly analysis-specific measure of distance or discrepancy.

There arises, then, a fundamental dilemma. On the one hand, a highly specific utility measure may yield a release tailored to a single analysis (or small class of related analyses), but that release may—unbeknownst to users—have low utility for other analyses. On the other hand, a broad utility measure may produce releases that are "pretty good" for a number of analyses, but "really good" for none. Worse yet, breadth seems almost invariably accompanied by bluntness: a broad measure may not be able to distinguish between quite different releases.

The principal purpose of this article is to construct a framework for thinking in a principled way about these kinds of issues, in the setting of numerical data. Inference-based measures for categorical data were discussed by Dobra et al. (2002) and Gomatam et al. (2005b). We illustrate our framework with:

- Two narrow measures that capture differences in the *inferences* based on $\mathcal{D}_{\text{rel}}$ and those based on $\mathcal{D}_{\text{orig}}$. As elaborated in Section 2.2, they are based on linear regression models for numerical data. (Their definability and relevance in broader settings are subjects of future research.)

- One broad measure—the Kullback–Liebler divergence $d_{\text{KL}}(\mathcal{D}_{\text{rel}}, \mathcal{D}_{\text{orig}})$ (Sec. 2.3).

In some ways, these could not be more different. The former is based on *one* particular model, with one designated response, but seeks to capture how inferences—not just point estimates of moments—relate. At the other extreme, $d_{\text{KL}}(\mathcal{D}_{\text{rel}}, \mathcal{D}_{\text{orig}})$ actually is a metric, so that (but only) in principle, if $d_{\text{KL}}(\mathcal{D}_{\text{rel}}, \mathcal{D}_{\text{orig}})$ is small, so should be all other reasonable measures of utility.

2.2 Narrow Measures

Data users often wish to fit linear regression models to numerical data. This process produces, of course, not only point estimates of the coefficients, but confidence intervals as well. Thus, it is clearly desirable to construct utility measures that indicate when the confidence interval based inferences from regressions using the released data are close to the corresponding ones using the original data.

We present two such measures. Although formulated for linear regressions, they can be extended, albeit not necessarily in a straightforward manner, to other analyses. These measures quantify the differences between inferences for one specific regression model, with the response and predictors designated *in advance* by the data disseminator. (This assumption is not as Draconian as it might seem initially. In many databases, there is one clearly identified response. Examples are education data in student performance is the response and epidemilogical studies in which survival time is the response.) How utilities for multiple models might be evaluated and combined is discussed further in Section 4.

2.2.1 *Confidence Interval Overlap*

Confidence intervals are the main mechanism of inference in regression models. Therefore, one measure of utility is the degree of overlap between confidence intervals obtained from the same regressions fit using the $\mathcal{D}_{\text{rel}}$ and $\mathcal{D}_{\text{orig}}$. The greater the overlap, the higher the utility.

Consider a prescribed regression, with specified response and predictors. Let $(L_{\text{rel},k}, U_{\text{rel},k})$ be the 95% confidence interval for the regression coefficient β_k obtained from $\mathcal{D}_{\text{rel}}$, and let $(L_{\text{orig},k}, U_{\text{orig},k})$ be the corresponding interval obtained from $\mathcal{D}_{\text{orig}}$. Let $f_{\text{rel},k}$ and $f_{\text{orig},k}$ be the estimated posterior distributions of β_k computed under $\mathcal{D}_{\text{rel}}$ and $\mathcal{D}_{\text{orig}}$, respectively. Specifically, $f_{\text{orig},k}$ is the usual t-distribution on $n-p$ degrees of freedom with mean $\hat{\beta}_{\text{orig},k}$ and variance the kth diagonal element in $\hat{\sigma}^2_{\text{orig}}\left(X'_{\text{orig}}X_{\text{orig}}\right)^{-1}$, where $\hat{\sigma}^2_{\text{orig}}$ is the estimated residual variance obtained from fitting the regression of Y_{orig} on the associated $n \times p$ matrix of predictors, X_{orig}, which includes a vector of ones for the intercept.

We define the probability overlap in the confidence intervals for any β_k to equal:

$$I_k = \frac{1}{2}\left[\int_{L_{\text{rel},k}}^{U_{\text{rel},k}} f_{\text{orig},k}(t)dt + \int_{L_{\text{orig},k}}^{U_{\text{orig},k}} f_{\text{rel},k}(t)dt\right] \qquad (3)$$

and the interval overlap measure IO as

$$I = \frac{1}{p}\sum_{i=1}^{p} I_k, \tag{4}$$

where p is the dimension of the predictor variable matrix, including the intercept.

By design, $0 \leq I_k \leq 0.95$ (as is the case for I), with effectively no overlap corresponding to $I_k = 0$ and perfect overlap corresponding to $I_k = 0.95$. Averaging the two integrals in the definition of I_k helps deal with cases where $(L_{\text{orig},k}, U_{\text{orig},k}) \subseteq (L_{\text{rel},k}, U_{\text{rel},k})$, or vice versa. For an illustrative example, consider the case where $(L_{\text{orig},k}, U_{\text{orig},k}) = (8, 10)$, and for two different proposed releases the $(L_{\text{rel}_1,k}, U_{\text{rel}_1,k}) = (-12, 30)$ and $(L_{\text{rel}_2,k}, U_{\text{rel}_2,k}) = (3, 15)$. From a utility perspective, the second release is clearly preferable over the first release. The IO as defined favors the second release. A criterion that just equals $\int_{L_{\text{rel},k}}^{U_{\text{rel},k}} f_{\text{orig},k}(t)dt$ does not clearly distinguish the releases, because this integral for both procedures is essentially one. Similar examples can be constructed to show the inadequacy of using $\int_{L_{\text{orig},k}}^{U_{\text{orig},k}} f_{\text{rel},k}(t)dt$ alone.

The IO does not distinguish among intervals that have I_k essentially equal to zero, some of which may be "less worse" than others. To adjust for this, the measure can be modified by adding some distance-based penalty when I is essentially zero, or perhaps even when I_k is essentially zero for some k, where distance is defined as some function of the $|\hat{\beta}_{\text{rel},k} - \hat{\beta}_{\text{orig},k}|$ or of $\min\{|L_{\text{rel},k} - U_{\text{orig},k}|, |L_{\text{orig},k} - U_{\text{rel},k}|\}$.

An alternative measure is the overlap in the interval lengths. Let $(L_{\text{over},k}, U_{\text{over},k})$ be the overlap in these intervals, defined as $\{b : b \geq L_{\text{orig},k}, b \geq L_{\text{rel},k}, b \leq U_{\text{orig},k}, b \leq U_{\text{rel},k}\}$. Then, the average relative overlap in the confidence intervals for any β_k equals

$$J_k = \frac{1}{2}\left[\frac{U_{\text{over},k} - L_{\text{over},k}}{U_{\text{orig},k} - L_{\text{orig},k}} + \frac{U_{\text{over},k} - L_{\text{over},k}}{U_{\text{rel},k} - L_{\text{rel},k}}\right]. \tag{5}$$

The interval overlap measure then could be defined as $J = (1/p)\sum_{i=1}^{p} J_k$.

2.2.2 *Ellipsoid Overlap*

The IO measure considers each interval separately, effectively using all the conditional distributions of the coefficients rather than their joint distribution. Some analysts may be interested in simultaneous intervals, which are defined by multidimensional ellipsoids. We therefore create an ellipsoid overlap measure, EO. Higher values of EO mean greater utility.

To construct EO it is convenient to consider posterior probabilities of regions defined by ellipsoids, that is, to use a Bayesian perspective. Generically, let $\hat{\beta}$ be the maximum likelihood estimate of β, the $p \times 1$ vector of true coefficients in the regression of Y on X, and let $\hat{\sigma}^2$ be the estimated residual variance for that regression. Under the standard linear regression assumptions and assuming standard noninformative prior distributions for β and σ^2, the $(1-\alpha)100\%$ joint highest posterior density ellipsoid for β is defined by all the values of β such that

$$\frac{(\beta - \hat{\beta})^T (X^T X)(\beta - \hat{\beta})}{p\hat{\sigma}^2} \leq F(\alpha; p, n-p),$$

where $F(\alpha; p, n-p)$ is the critical value from the F distribution with p and $n-p$ degrees of freedom. The ellipsoid from the $\mathcal{D}_{\text{orig}}$, which we call E_{orig}, is obtained by setting $\hat{\beta} = \hat{\beta}_{\text{orig}}$, $\hat{\sigma}^2 = \hat{\sigma}^2_{\text{orig}}$, and $X = X_{\text{orig}}$. The ellipsoid from the $\mathcal{D}_{\text{rel}}$, which we call E_{rel}, is obtained by setting $\hat{\beta} = \hat{\beta}_{\text{rel}}$, $\hat{\sigma}^2 = \hat{\sigma}^2_{\text{rel}}$, and $X = X_{\text{rel}}$.

The utility measure EO is the average of two posterior probabilities: (1) the probability of E_{orig} computed using the posterior distribution of β based on $\mathcal{D}_{\text{rel}}$, and (2) the probability of E_{rel} computed using the posterior distribution of β based on $\mathcal{D}_{\text{orig}}$. To determine these probabilities, we use Monte Carlo simulations. For the first probability, we draw values of β from its posterior conditional on $\mathcal{D}_{\text{rel}}$ which is a p-variate t-distribution with mean $\hat{\beta}_{\text{rel}}$ and covariance matrix $\hat{\Sigma}_{\text{rel}} = \hat{\sigma}^2_{\text{rel}}(X^t_{\text{rel}} X_{\text{rel}})^{-1}$ with $n-p$ degrees of freedom. We then calculate the percentage of these drawn β that lie within E_{orig}. A similar process is used to obtain the second probability by drawing from the posterior of β given $\mathcal{D}_{\text{orig}}$ and finding the percentage of these that lie inside E_{rel}. As with IO, the EO can be extended to any parameters whose distribution is well-approximated by a multivariate normal distribution.

2.3 Broad Measures

At the opposite end of the utility spectrum, one can employ broad measures of the overall difference between $\mathcal{D}_{\text{rel}}$ and $\mathcal{D}_{\text{orig}}$, of which the broadest are metrics on some set of distributions.

This article focuses on the Kullback-Liebler divergence between (the empirical distribution of) $\mathcal{D}_{\text{rel}}$ and that of $\mathcal{D}_{\text{orig}}$, which we denote by $d_{\text{KL}}(\mathcal{D}_{\text{rel}}, \mathcal{D}_{\text{orig}})$. Since $\mathcal{D}_{\text{rel}}$ and $\mathcal{D}_{\text{orig}}$ are discrete distributions, calculation of $d_{\text{KL}}(\mathcal{D}_{\text{rel}}, \mathcal{D}_{\text{orig}})$ entails two computationally onerous steps:

1. Construction of density estimators $\hat{f}_{\text{rel}}$ and $\hat{f}_{\text{orig}}$.
2. Approximation of

$$d_{\text{KL}}(\mathcal{D}_{\text{rel}}, \mathcal{D}_{\text{orig}}) = \int \log\left[\hat{f}_{\text{rel}}/\hat{f}_{\text{orig}}\right]\hat{f}_{\text{rel}} \tag{6}$$

by numerical quadrature.

In high (in practice, three or more) dimensions, both of these may be infeasible.

When both $\mathcal{D}_{\text{rel}}$ and $\mathcal{D}_{\text{orig}}$ have multivariate normal distributions, $d_{\text{KL}}(\mathcal{D}_{\text{rel}}, \mathcal{D}_{\text{orig}})$ can be calculated in closed form. The resultant expression (A.4), which is used in Section 3.3, is derived in the Appendix.

3. ILLUSTRATIVE APPLICATIONS OF THE UTILITY FRAMEWORK

This section presents two applications of the utility framework. The first illustrates the risk-utility framework using "real data" from the Current Population Survey (CPS). The second uses a simulation study to explore the properties of utility measures and SDL procedures as a function of the size and correlation structure of the original data.

In both applications, we use a representative set of SDL methods investigated by Oganian (2003), which are described in

Section 3.1. This should not be construed as endorsing these methods—indeed, one of them seems to have rather undesirable properties, nor should it be construed as disparaging other methods.

To measure disclosure risk, following the example of Yancey, Winkler, and Creecy (2002), we determine the percentage of records in $\mathcal{D}_{\text{rel}}$ that we can match correctly to records in $\mathcal{D}_{\text{orig}}$ using standard record linkage techniques (Felligi and Sunter 1969; Jaro 1989). For simplicity, we do not consider other measures of identity disclosure risk, nor measures of attribute disclosure risk, although we believe that data disseminators should consider such measures.

3.1 Disclosure Limitation Methods

In a taxonomy of SDL methods that release microdata, the highest level distinction is whether they are record-level or database-level. For record-level methods, the released data are

$$\mathcal{D}_{\text{rel}} = \{f(r) : r \in \mathcal{D}_{\text{orig}}\}, \qquad (7)$$

where r is a record in $\mathcal{D}_{\text{orig}}$ and f is a function that does not depend on $\mathcal{D}_{\text{orig}}$, but may involve exogenous randomness. That is, records are simply altered individually, for example, by addition of noise. Database-level methods are more complex: in effect, the function f in (7) is replaced by $f(\mathcal{D}^0_{\text{orig}}(r))$, where $\mathcal{D}^0_{\text{orig}}(r)$ is a subset of $\mathcal{D}_{\text{orig}}$ that in general depends on r and often involves exogenous randomness. Microaggregation and data swapping are of this nature. In the extreme case of synthetic data (Raghunathan et al. 2003; Reiter 2005b), $\mathcal{D}^0_{\text{orig}}(r) = \mathcal{D}_{\text{orig}}$ for all r. We consider both record-level and database level methods.

Virtually all SDL methods can be implemented with differing degrees of intensity. For example, one can add large or small amounts of noise to data. Hence, we write each SDL method as a function of the parameter that can be varied. Because our purpose is to illustrate the utility measures framework, in our experiments we select only one value of the parameter for each method. In future work, we plan to use the risk-utility framework to assess the sensitivity of SDL procedures to different parameter values.

3.1.1 *Additive Noise*

Additive noise (Brand 2002; Duncan and Pearson 1991; Kim 1986; Little 1993; Sullivan and Fuller 1989; Tendik and Matloff 1994) consists of adding random noise to the original data. Generally, the noise distribution has mean zero, to preserve, on average, the sample means. The variance of noise distribution can be generic, although most commonly it reflects either complete independence or the correlation structure of the original data.

In Sections 3.2 and 3.3, we employ Gaussian noise with the same correlation structure as the original data. Specifically, let $\mathbf{X}$ be original multivariate dataset with covariance matrix Σ_{orig}. The corresponding masked data are generated as

$$\mathbf{X}' = \mathbf{X} + \mathbf{E} \qquad (8)$$

$$\mathbf{E} \sim N(\mathbf{0}, c\Sigma_{\text{orig}}) \qquad (9)$$

where the constant c is defined by the data disseminator. When adding noise with the same correlation structure as $\mathcal{D}_{\text{orig}}$, the c is the parameter that defines the procedure. We set $c = 0.16$ in the simulations. We abbreviate this SDL method as `Noise(.16)`.

3.1.2 *Rank Swapping*

Rank swapping is a form of data swapping (Dalenius and Reiss 1982). It was originally designed for ordinal variables (Moore 1996), but works equally for numerical variables. To implement rank swapping, we first rank the values of variable X_i in ascending order. Each ranked value then is swapped with another ranked value randomly chosen within a restricted range. This process is repeated for each variable.

Typically, the swaps are defined by setting a parameter p so that the ranks of two swapped values are not allowed to differ by more then p percent of the total number of records. In the example above, $p = 10\%$ corresponding to swapping with the next ordered value. Large values of p lead to greater distortions in the data whereas the smaller ones to higher disclosure risk. In Domingo-Ferrer and Torra (2001), Oganian (2003), and Domingo-Ferrer et al. (2001), $p = 15\%$ was reported as one of the best parameter choices for rank swapping. We therefore used this parameter value in our simulations. We abbreviate this method as `Rank(.15)`.

3.1.3 *Microaggregation*

Microaggregation involves clustering records into small aggregates or groups of size at least k. Rather than releasing the original value of X_i for a given record, the disseminator releases the average of the original values of X_i for a group of records. Classical microaggregation requires that all groups, except perhaps one, be of size k, where k is selected by the data disseminator (Defays and Nanopoulos 1993).

We examined several variants of microaggregation in our simulations, each a function of which and how many variables and records are grouped together. These include: (1) individual ranking, in which each variable is grouped independently of other variables; (2) multivariate ranking, in which the variables are grouped by similarity of values for subsets of variables; and (3) z-scores projection and principal components projection (Anwar 1993; Defays and Nanopoulos 1993; Defays and Anwar 1995), in which the multivariate data first are ranked by projecting them onto a single axis, using either the sum of z-scores or the first principal component, and then are aggregated into groups of size k, except possibly for one group of larger size (from $k+1$ to $2k-1$).

Microaggregation methods are functions of the number of variables used in the similarity measures (v), and the group sizes (k). We set values for v and p according to the research done by Domingo-Ferrer and Torra (2001) and Oganian (2003). For individual ranking, we used all variables in the similarity measures ($v = p$) and ten records per group ($k = 10$). This method is abbreviated as `Micir(p,10)`. For multivariate ranking, we considered several approaches. First, we used all variables in the similarity measures and three records per group. This method is abbreviated as `Micm(p,3)`. Second, we used three variables at a time in the similarity measures—for example, replace variables X_1 through X_3 with a group average, then replace variables X_4 through X_6 with an independently formed group average, and

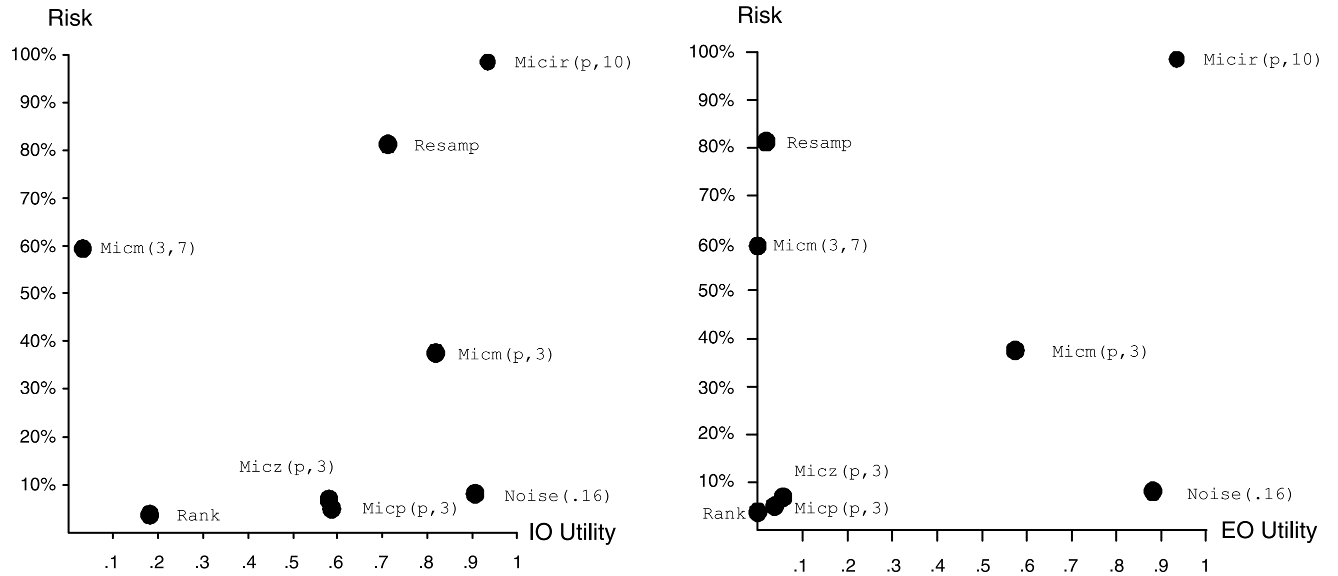

Figure 1. Risk-utility plots for the 1995 CPS data. Left: IO measure. Right: EO measure. Higher values of IO and EO represent greater utility.

so on—and seven records per group. This method is abbreviated as `Micm(3,7)`. Finally, for both forms of microaggregation on projected data, we used all variables in the projection scores and three records per group. These are abbreviated as `Micp(p,3)` for principal components projection and `Micz(p,3)` for z-scores projection.

3.1.4 Resampling

Resampling is a generic term, but here we mean a specific approach to protecting data that involves elements of bootstrapping. This version was used by Domingo-Ferrer and Mateo-Sanz (1999) and Heer (1993). Let X_1 be the first variable in a dataset with n records. We give each row a ranking based on its value of X_1, which is determined by its position in an ascending sort of X_1. We then draw n values from the data in X_1, with replacement, and order them consistent with the ordering of the row ranks to obtain a bootstrap sample V_{11}. This process is repeated independently t times, resulting in bootstrap samples $V_{11}, \ldots, V_{1t}$. The released X_1 is $\bar{V}_1 = (1/t)\sum_{k=1}^{t} V_{1k}$. We repeat this process independently for each X_i, for $i = 1, \ldots, p$, by ranking the rows in ascending order of the X_i and bootstrapping to obtain $V_{i1}, \ldots, V_{it}$. The released dataset is $(\bar{V}_1, \bar{V}_1, \ldots, \bar{V}_p)$.

For resampling, the parameter is t, the number of bootstrap samples, and we use $t = 3$. This method is abbreviated as `Resamp(3)`.

3.2 Application 1: Risk-Utility Tradeoffs on CPS Data

Utility measures must be assessed in combination with disclosure risk measures to quantify the risk-utility tradeoffs of various SDL procedures. Here, we illustrate such quantifications using microdata extracted from the 1995 CPS. The data comprise 1080 records containing 12 numerical variables, including adjusted gross income (`agi`), employer contribution for health insurance (`emcontrb`), business or farm net earnings (`ernval`), federal income tax liability (`fedtax`), social security retirement payroll reduction (`FICA`), amount of interest income (`intval`), total person earnings (`pearnval`), total other persons income (`pothval`), total person income (`ptotval`), state income tax liability (`statetax`), taxable income amount (`taxinc`), and total wage and salary (`wsalval`). These variables are highly correlated; in fact, the income variables contain a perfect linear combination.

We quantify disclosure risk as the percentage of records in $\mathcal{D}_{\text{rel}}$ that can be linked correctly to their "parent" records in $\mathcal{D}_{\text{orig}}$, assuming that the intruder knows the exact values for six variables in the data set—`fedtax`, `agi`, `emcontrb`, `ptotval`, `taxinc`, and `statetax`, and that these values equal the corresponding values in $\mathcal{D}_{\text{orig}}$. These six were chosen because each alone uniquely identified all individuals in the dataset, so that they are the "riskiest" set of six variables one could know in these data. In general, data disseminators can assess disclosure risk under a variety of assumptions about intruders' knowledge, as was done for example by Fienberg et al. (1997) and Reiter (2005a).

For the model-specific utility measures, the regression of interest is

$$\begin{aligned} \texttt{agi} &= \beta_0 + \beta_1\,\texttt{emcontrb} + \beta_2\,\texttt{fedtax} \\ &\quad + \beta_3\,\texttt{taxinc} + \beta_4\,\texttt{ptotval} + \beta_5\,\texttt{statetax} + \varepsilon. \qquad (10) \end{aligned}$$

We fit the regression using both $\mathcal{D}_{\text{orig}}$ and the $\mathcal{D}_{\text{rel}}$ resulting from the various SDL strategies.

Figure 1 displays (risk, utility) scatterplots of the values of IO and EO (x-axis) and disclosure risk (y-axis) for each of the SDL strategies in Section 3.1. We do not calculate the Kullback-Liebler divergence $d_{\text{KL}}(\mathcal{D}_{\text{rel}}, \mathcal{D}_{\text{orig}})$ of (6) because $\mathcal{D}_{\text{orig}}$ does not follow a multivariate normal distribution. In all cases, EO $\leq$ IO; for some measures the drop is precipitous.

The risk-utility frontiers associated with (2), in order of decreasing utility, are:

For IO, `Micz(p,3)`, `Noise(.16)`, `Micp(p,3)`, `Rank(.15)`.

For EO, `Micz(p,3)`, `Noise(.16)`, `Micz(p,3)`, `Micp(p,3)`, `Rank(.15)`.

Not surprisingly, the former is a subset of the latter. The data disseminator can ignore `Micm(3,7)`, `Micm(p,3)` and `Resamp(3)` for both utility measures.

The choice among the SDL methods lying on the risk-utility frontier lies with the data disseminator. To illustrate the first approach described in Section 1, if the risk threshold were 10% (in some settings, not a very conservative value), then according to either IO or EO, `Noise(.16)` would be the preferred SDL method. It is also clear from Figure 1 that compared to `Micz(p,3)` or `Noise(.16)`, `Micir(p,10)` produces only a minor increase in utility at an enormous cost in terms of disclosure risk. Similarly, `Rank(.15)` yields only a modest improvement in disclosure risk over `Micp(p,3)` and `Noise(.16)`, but incurs an immense penalty in terms of data utility, especially for EO. Thus, it appears that in practice the disseminator would choose `Noise(.16)` or `Micp(p,3)` for the model in (10).

How data disseminators might examine multiple analyses and multiple SDL strategies in the process of selecting $\mathcal{D}_{\text{rel}}$ is discussed in Section 4.

3.3 Application 2: Properties of SDL Procedures and Utility Measures

In Figure 1, there is no clear difference between IO and EO. Moreover, the one case study in Section 3.2, which involves only one database, does yield insight into how characteristics of $\mathcal{D}_{\text{orig}}$ might affect utility measures and consequent choice of SDL methods. This section reports simulation studies designed to provide answers to these kinds of questions.

The design for the simulation consists of:

- Six data types constructed by crossing two correlation structures—high and low—with three dimension structures—three, six, and ten variables. Each simulated data set comprises 10,000 observations drawn from a multivariate normal distribution.

- Five replicates for each data type, to assess the effects of replicate variability.

- The eight SDL measures from Section 3.1.

- The two narrow utility measures from Section 2.2 and the one broad measure from Section 2.3. For the model-specific utility measures, we selected one variable as the response (this variable is present regardless of the dimension) and regressed it on all other (2, 5, or 9) variables in the dataset.

Disclosure risk is the percentage of records identified correctly using record linkage on all variables in the dataset.

For the Kullback–Liebler divergence measure KL $= d_{\text{KL}}(\mathcal{D}_{\text{rel}}, \mathcal{D}_{\text{orig}})$, we assume that $\mathcal{D}_{\text{rel}}$ has a multivariate normal distribution ($\mathcal{D}_{\text{orig}}$ has one by construction). Equation (A.4) in the appendix was used to calculate KL, using maximum likelihood estimators $(\hat{\mu}_{\text{orig}}, \hat{\Sigma}_{\text{orig}})$ for the mean and covariance of $\mathcal{D}_{\text{orig}}$ and $(\hat{\mu}_{\text{rel}}, \hat{\Sigma}_{\text{rel}})$ for the mean and covariance of $\mathcal{D}_{\text{rel}}$. The assumption that $\mathcal{D}_{\text{rel}}$ is multivariate normal is an approximation at best. In what follows, it is important to keep in mind that smaller values of KL indicate higher utility.

Table 1. Risk and Utility Values in Simulated Low-Correlation, Multivariate Normal Data. Values in boldface are on the risk-utility frontier.

Method	*Dim*	*EO*	*IO*	*KL*	*Risk*
`Micir(p,10)`	3	**0.949**	**0.950**	**3.93E-07**	0.948
	6	**0.950**	**0.950**	**2.36E-06**	0.974
	10	**0.945**	**0.948**	**3.47E-05**	0.985
`Resamp(3)`	3	0.780	0.916	1.71E-04	0.455
	6	0.622	0.843	**0.001**	0.735
	10	0.106	0.867	**0.004**	0.846
`Micp(p,3)`	3	0.000	1.87E-20	0.902	0.018
	6	0.000	0.038	2.237	0.030
	10	0.000	0.614	**4.067**	0.057
`Rank(.15)`	3	**0.000**	**2.16E-12**	**0.081**	0.001
	6	**0**	**6.01E-05**	**0.334**	0.004
	10	0	0.156	**0.987**	0.066
`Micm(3,7)`	3	0.761	0.83	**0.001**	0.110
	6	0.008	0.644	**0.010**	0.245
	10	0.000	0.666	0.287	0.550
`Micm(p,3)`	3	**0.930**	**0.933**	**1.55E-04**	0.120
	6	2.78E-05	0.423	0.080	0.141
	10	0.134	0.738	0.449	0.238
`Micz(p,3)`	3	0	0.0	0.903	0.005
	6	0	**0.219**	2.260	0.005
	10	**0.15**	**0.755**	**4.129**	0.009
`Noise(.16)`	3	**0.916**	**0.926**	**0.016**	0.003
	6	**0.907**	**0.929**	**0.031**	0.017
	10	**0.870**	**0.920**	**0.053**	0.108

Tables 1 and 2 present the utility and risk values, averaged over replicates, for the low- and high-correlation datasets. The standard errors of the reported averages are all small enough that observed differences do not result solely from replicate variability in the simulations. Boldface utility values indicate that the SDL procedure is on the risk-utility frontier for that utility measure; that is, the procedure is not dominated by other procedures.

Looking at the Tables 1 and 2, differences in risk and utility across methods are larger than differences due to either dimension or correlation structure. `Micir(p,10)` typically provides the highest utilities but also the highest disclosure risks, which is consistent with the results in Section 3.2. (And not surprising—this method usually alters the observed data only slightly.) `Resamp(3)` tends to have the second highest disclosure risk, with relatively high utility. At the "low end," `Rank(.15)` typically has among the lowest disclosure risks and the lowest utilities, because it alters significantly the correlation structure of the data, greatly distorting regression inferences. `Micm(3,7)`, which does microaggregation on three variables at a time, tends to have high disclosure risk and low utility, due to independent aggregations of different triplets of variables. Among the microaggregation methods that operate on all variables simultaneously—`Micm(p,3)`, `Micp(p,3)`, and `Micz(p,3)`—`Micm(p,3)` generally has highest utility, especially for the KL measure, and highest risk. `Noise(.16)` is characterized by relatively high utility and low disclosure risk; it is the only method that purposefully

Table 2. Risk and Utility Values in Simulated High-Correlation, Multivariate Normal Data. Values in boldface are on the risk-utility frontier.

Method	*Dim*	**EO**	**IO**	**KL**	*Risk*
`Micir(p,10)`	3	**0.946**	**0.948**	**4.65E-06**	0.947
	6	0.834	0.760	**0.003**	0.972
	10	0.763	0.735	**0.027**	0.985
`Resamp(3)`	3	0.364	0.883	0.001	0.402
	6	0.000	0.873	**0.018**	0.664
	10	0.000	0.487	0.118	0.833
`Micp(p,3)`	3	0.000	0.428	0.888	0.035
	6	0.000	0.602	2.264	0.034
	10	0.108	0.811	4.051	0.043
`Rank(.15)`	3	**0.000**	**9.23E-15**	**0.720**	0
	6	**0.000**	**0.034**	**2.887**	0.001
	10	**0.000**	**0.097**	**5.908**	0.004
`Micm(3,7)`	3	0.739	0.706	**0.004**	0.150
	6	0.000	0.150	0.387	0.155
	10	0.000	0.118	1.736	0.419
`Micm(p,3)`	3	**0.923**	0.912	**0.001**	0.161
	6	0.000	0.443	0.512	0.181
	10	0.539	0.843	1.359	0.281
`Micz(p,3)`	3	0.000	0.694	0.930	0.015
	6	0.306	0.846	2.267	0.021
	10	**0.367**	**0.750**	**4.072**	0.032
`Noise(.16)`	3	**0.920**	**0.930**	**0.016**	0.002
	6	**0.871**	**0.921**	**0.031**	0.011
	10	**0.827**	**0.904**	**0.053**	0.040

preserves the correlation structure of the data.

The first three columns of Table 3 display, for each of the three utility measures, the number of times—out of a possible six corresponding to the six database structures—each method is on the risk-utility frontier.

The corresponding column totals indicate that the frontiers for the model-specific measures are smaller than the frontier for KL. That is, more methods are dominated by others when using IO and EO. The results highlight `Noise(.16)`, `Micir(p,10)`, `Micz(p,3)`, and `Rank(.15)` as being on the frontier most often. As shown in Tables 1 and 2, as well as in Section 3.2, `Micir(p,10)` and `Rank(.15)` tend to be at the extreme ends of the utility or risk portion of the frontier, whereas `Noise(.16)` and `Micz(p,3)` lie in the middle of the frontier.

One advantage of the risk-utility frontier formulation in (2) is that it extends to more than one utility measure (or more than one measure of risk). If there are multiple utility measures $\mathrm{DU}_1, \ldots, \mathrm{DU}_k$, then the partial order is defined by

$$R_1 \preceq_{\mathrm{RU}} R_2 \Leftrightarrow \mathrm{DR}(R_2) \quad \leq \quad \mathrm{DR}(R_1)$$

and

$$\mathrm{DU}_i(R_2) \quad \geq \quad \mathrm{DU}_i(R_1) \quad \text{for} \quad i = 1, \ldots, k. \tag{11}$$

Of course, as the number of measures increases, so does the relative size of the frontier, reducing the savings from restricting attention to the frontier.

The fourth and fifth columns of Table 3 show how many times each method is on the joint frontier for IO and EO and how many times on the joint frontier for all three utility measures. The joint {IO, EO} frontier is reasonably close to the individual frontiers, while the three-measure frontier is quite different, especially for procedures `Micm(p,3)` and `Resamp(3)`. This reflects the low discriminatory power of KL.

Tables 1 and 2 also provide some insight into the effects of dimension and correlation structure. For several methods, the value of EO is essentially zero, indicating little probability mass in the intersections of the ellipsoids. Because EO measures simultaneous overlap, any substantial disparity between the distributions of the parameters, even in just one dimension, results in a low value of EO. This issue also explains why values of EO tend to decrease as dimension increases: there are more opportunities for disparities, and small disparities add up to produce bigger joint differences. A similar behavior applies for the KL measure. In contrast, the IO measure rarely equals zero, and there is no strong dimension effect. This is because IO averages individual overlaps, so that overlap in several dimensions contribute positive values even when one dimension is poorly specified.

In general, the differences in the confidence intervals based on $\mathcal{D}_{\text{orig}}$ and $\mathcal{D}_{\text{rel}}$ are larger in the high-correlation data than in the low-correlation data, but this effect is weak relative to that of differing methods. Some methods, such as `Noise(.16)`, seem to be essentially unaffected by the correlation structure, whereas others, such as `Micz(p,3)`, are strongly affected. Among the utility measures, KL appears most sensitive to the correlation structure, especially for `Rank(.15)` and some of the microaggregation methods.

We also examined the performance of the methods when the analyst fits an incorrect model—one that excludes important predictors—but have omitted detailed numerical results. Some of the SDL methods produced regressions bearing little resemblance to the corresponding regressions fit with the original data. This was especially true for some of the microaggregation methods, which should give pause to disseminators considering use of microaggregation. The finding also emphasizes the importance of checking several inferences when doing risk-utility analyses.

Table 3. Numbers of Times Each SDL Method Appears on the Marginal and Joint Risk-Utility Frontiers for the Six Simulated Datasets

Methods	*EO*	*IO*	*KL*	*Joint EO, IO*	*Joint EO, IO, KL*	*Total*
`Noise(.16)`	6	6	6	6	6	30
`Rank(.15)`	5	5	6	5	6	27
`Micir(p,10)`	4	4	6	4	6	24
`Micz(p,3)`	2	3	2	3	3	13
`Micm(p,3)`	2	1	2	2	2	9
`Micm(3,7)`	0	0	3	0	3	6
`Resamp(3)`	0	0	3	0	3	6
`Micp(p,3)`	0	0	1	0	1	2
Total	21	20	31	20	30	

4. DISCUSSION

As threats to data confidentiality grow, agencies and survey organizations must implement disclosure limitation with increasing intensity. Deciding which procedures to use, as well as how intensely to use them, can—and, we would argue, should—be framed in the context of a risk-utility analysis. The utility measures presented here can aid in quantifying that tradeoff.

These measures have strengths and weaknesses. The interval and ellipse overlap measures can be used for many types of inferences, but they are specific not just to a class of models, but to one model within a class. One of the attractive features of public use data releases is that a variety of analyses can be performed on them. This makes it infeasible to predict all inferences that will be attempted, but clearly certain inferences can be identified as more typical, and hence more important to preserve, than others. For example, predicting income from age is more typical than predicting age from income.

When multiple models are of interest, one approach is to employ multidimensional utilities (as, albeit in a different context, in Section 3.3), and to define risk-utility frontiers using analogs of (11). In this case, there is one utility measure per model of interest. When there are many models of interest, this approach is cumbersome at best, since most or nearly all candidate releases may be on the frontier (especially if both IO and EO are to be considered).

An alternative is to use a loss function to combine model-specific utilities for a large number of representative models that have been identified from existing literature and subject matter expertise. For instance, a weighted linear combination of model-specific utilities could be used, as in the experimental design literature, where design points can be selected to optimize for a set of linear regressions.

Indeed, as suggested by a referee of this article, there may be deeper connections between microdata release and experimental design. To illustrate, random sampling (of records) from a database is a common SDL strategy because it increases intruder uncertainty about whether a target record is present in the released data. While choice of a design matrix, as in Chaloner (1984), has no direct parallel in SDL, one intriguing analog would be to release a set of records that preserve fidelity of a family of regressions. Doing so produces a formulation very similar to ψ-optimality in Chaloner (1984). However, instead of the unconstrained optimization problem there, one faces a daunting discrete optimization problem because the "design" must be selected from the underlying database. Of course, the disclosure risk consequences of such a strategy are completely unclear.

These kinds of approaches, methods of selecting representative analyses, and useful and workable tools for combining model-specific utilities are topics for future research.

We investigated one broad measure, KL, but it relies on the multivariate normality assumption to be meaningful. Data disseminators would benefit greatly from the development of computationally feasible techniques to measure distances between empirical distributions.

Finally, it may be important for data disseminators to evaluate relationship-specific measures of utility, although we did not illustrate them here. One such measure is the number of substantively important, statistically significant relationships that experience a directional switch, for example, the estimated regression coefficient goes from positive to negative, when going from $\mathcal{D}_{\text{orig}}$ to $\mathcal{D}_{\text{rel}}$. Clearly, a release that involves many directional switches is undesirable. The rationale is that a change in sign mis-states the direction of an effect. A related measure is the number of relationships that go from statistically significant to statistically insignificant, or vice versa: many significance changes are undesirable from a utility perspective. These relationship measures complement the model-specific measures in the utility evaluation process.

APPENDIX: DERIVATION OF THE KULLBACK-LIEBLER DIVERGENCE FOR MULTIVARIATE NORMAL DISTRIBUTIONS

Let X_1 and X_2 be p-dimensional random variables with multivariate normal densities $\phi_1 = \mathrm{MVN}(\mu_1, \Sigma_1)$ and $\phi_2 = \mathrm{MVN}(\mu_2, \Sigma_2)$. Then by (6),

$$\begin{aligned} d_{\mathrm{KL}}(X_1 \| X_2) &= E_{X_1}\Big[\frac{1}{2}\log\left(|\Sigma_2|/|\Sigma_1|\right) \\ &\quad -\frac{1}{2}\left[(X-\mu_1)'\Sigma_1^{-1}(X-\mu_1)\right. \\ &\quad \left. -(X-\mu_2)'\Sigma_2^{-1}(X-\mu_2)\right]\Big] \\ &= \frac{1}{2}\log\left(|\Sigma_2|/|\Sigma_1|\right) - \frac{1}{2}E_{X_1}\left[T_1\right] \\ &\quad +\frac{1}{2}E_{X_1}\left[T_2\right], \end{aligned}$$

where $T_1 = (X-\mu_1)'\Sigma_1^{-1}(X-\mu_1)$ and $T_2 = (X-\mu_2)'\Sigma_2^{-1}(X-\mu_2)$. Under the distribution of X_1, $T_1 \sim \chi^2_p$, so that

$$E_{X_1}[T_1] = p. \tag{A.1}$$

Also, we can re-express T_2 as

$$\begin{aligned} T_2 &= (X-\mu_2)'\Sigma_2^{-1}(X-\mu_2) \\ &= (X-\mu_1)'\Sigma_2^{-1}(X-\mu_1) \\ &\quad +2X'\Sigma_2^{-1}(\mu_1-\mu_2) - \mu_1\Sigma_2^{-1}\mu_1 + \mu_2\Sigma_2^{-1}\mu_2 \\ &= (X-\mu_1)'\Sigma_2^{-1}(X-\mu_1) + (\mu_1-\mu_2)'\Sigma_2^{-1}(\mu_1-\mu_2). \end{aligned} \tag{A.2}$$

Under the distribution of X_1 the quadratic form $(X-\mu_1)'\Sigma_2^{-1}(X-\mu_1)$ has a weighted χ^2 distribution of the form $\sum_{i=1}^p \lambda_i \chi^2_1$, where λ_i are the eigenvalues of $\Sigma_1\Sigma_2^{-1}$ (Guttman 1982). Hence,

$$E_{X_1}[T_2] = \sum_{i=1}^p \lambda_i + (\mu_1-\mu_2)'\Sigma_2^{-1}(\mu_1-\mu_2). \tag{A.3}$$

After noting that $\frac{1}{2}\log\left(|\Sigma_2|/|\Sigma_1|\right) = -\sum_{i=1}^p \log(\lambda_i)$, we obtain from (A.1) and (A.3) that

$$d_{\mathrm{KL}}(X_1 \| X_2) = \frac{1}{2}\left[(\mu_1-\mu_2)'\Sigma_2^{-1}(\mu_1-\mu_2) - \sum_{i=1}^p (1 - \lambda_i + \log[\lambda_i])\right]. \tag{A.4}$$

[Received TKK. Revised TKKK.]

REFERENCES

Anwar, N. (1993). "Micro-Aggregation—The Small Aggregates Method," Research Report.

Brand, R. (2002), "Microdata Protection Through Noise," in *Inference Control in Statistical Databases* (vol. 2316), eds. J. Domingo-Ferrer, Berlin: Springer–Verlag, Berlin, pp. 97–116.

Chaloner, K. (1984), "Optimal Bayesian Experimental Design for Linear Models," *The Annals of Statistics*, 12, 283–300.

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

Defays, D., and Anwar, N. (1995), "Micro-Aggregation: A Generic Method," in *Proceedings of the 2nd International Symposium on Statistical Confidentiality*, Luxembourg: Office for Official Publications of the European Community, pp. 69–78,

Defays, D., and Nanopoulos, P. (1993), "Panels of Enterprises and Confidentiality: The Small Aggregates Method," in *Proceedings of the 92 Symposium on Design and Analysis of Longitudinal Surveys*, Ottawa: Statistics Canada, pp. 195–204.

Dobra, A., Fienberg, S. E., Karr, A. F., and Sanil, A. P. (2002), "Software Systems for Tabular Data Releases," *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems*, 10, 529–544.

Dobra, A., Karr, A. F., and Sanil, A. P. (2003), "Preserving Confidentiality of High-Dimensional Tabular Data: Statistical and Computational Issues," *Statistics and Computing*, 13, 363–370.

Domingo-Ferrer, J., and Mateo-Sanz, J. M. (1999), "On Resampling for Statistical Confidentiality in Contingency Tables," *Computers and Mathematics with Applications*, 38, 13–32.

Domingo-Ferrer, J., Mateo-Sanz, J. M., and Torra, V. (2001), "Comparing SDC Methods for Microdata on the Basis of Information Loss and Disclosure Risk," in *Proceedings of the ETK–NTTS 2001*, Luxembourg: Eurostat, pp. 807–825.

Domingo-Ferrer, J., and Torra, V. (2001), "A Quantitative Comparison of Disclosure Control Methods for Microdata," in *Confidentiality, Disclosure and Data Access*, eds. P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz, Amsterdam: North–Holland, pp. 111–133.

Duncan, G. T., and Fienberg, S. E. (1999), "Obtaining Information While Preserving Privacy: A Markov Perturbation Method for Tabular Data," in *Eurostat Statistical Data Protection '98 Lisbon*, Luxembourg: Eurostat, pp. 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*, eds. P. Doyle, J. I. Lane, J. J. M. Theeuwes, and L. V. Zayatz, Amsterdam: Elsevier, pp. 135–166.

Duncan, G. T., Keller-McNulty, S. A., and Stokes, S. L. (2004), "Disclosure Risk vs. Data Utility: The R–U Confidentiality Map," *Management Science*, submitted for publication.

Duncan, G. T., and Lambert, D. (1986), "Disclosure-Limited Data Dissemination," *Journal of the American Statistical Association*, 81, 10–28.

——— (1989), "The Risk of Disclosure for Microdata," *Journal of Business and Economic Statistics*, 7, 207–217.

Duncan, G. T., and Pearson, R. W. (1991), "Enhancing Access to Microdata While Protecting Confidentiality: Prospects for the Future," *Statistical Science*, 6, 219–239.

Felligi, I. P., and Sunter, A. B. (1969), "A Theory for Record Linkage," *Journal of the American Statistical Association*, 64, 1183–1210.

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

Gomatam, S., Karr, A. F., Reiter, J. P., and Sanil, A. P. (2005a), "Data Dissemination and Disclosure Limitation in a World Without Microdata: A Risk-Utility Framework for Remote Access Analysis Servers," *Statistical Science*, 20, 163–177.

Gomatam, S., Karr, A. F., and Sanil, A. P. (2005b), "Data Swapping as a Decision Problem," *Journal of Official Statistics*, to appear. Available online at *www.niss.org/dgii/technicalreports.html*.

Guttman, I. (1982), *Linear Models: An Introduction*, New York: Wiley.

Heer, G. R. (1993), "A Bootstrap Procedure to Preserve Statistical Confidentiality in Contingency Tables," in *Proceedings of the International Seminar on Statistical Confidentiality*, ed. D. Lievesley, Luxembourg: Office for Official Publications of the European Community, pp. 261–271.

Jaro, M. A. (1989), "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida," *Journal of the American Statistical Association*, 84, 414–420.

Karr, A. F., Sanil, A. P., and Banks, D. L. (2006), "Data Quality: A Statistical Perspective," *Statistical Methodology*, 3, 137–173.

Kim, J. J. (1986), "A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation," in *Proceedings of the ASA Section on Survey Research Methodology*, Alexandria VA: American Statistical Association, pp. 303–308.

Kung, H. T., Luccio, F., and Preparata, F. P. (1975), "On Finding the Maxima of a Set of Vectors," *Journal of the ACM*, 22, 469–476.

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.

Moore, R. (1996), "Controlled Data Swapping Techniques for Masking Public use Microdata Sets," U. S. Census Bureau.

Oganian, A. (2003), "Security and Information Loss in Statistical Database Protection," Ph.D. thesis, Universitat Politècnica de Catalunya.

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

Reiter, J. P. (2005a), "Estimating Identification Risks in Microdata," *Journal of the American Statistical Association*, 100, 1101–1113.

——— (2005b), "Releasing Multiply-Imputed, Synthetic Public use Microdata: An Illustration and Empirical Study," *Journal of the Royal Statistical Society*, Series A, 168, 185–205.

Skinner, C. J., and Elliot, M. J. (2002), "A Measure of Disclosure Risk for Microdata," *Journal of the Royal Statistical Society*, Series B, 64, 855–867.

Sullivan, G., and Fuller, W. A. (1989), "The Use of Measurement Error to Avoid Disclosure," in *Proceedings of the ASA Section on Survey Research Methodology*, Alexandria, VA: American Statistical Association, pp. 802–807.

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

Trottini, M. (2003), "Decision Models for Data Disclosure Limitation," unpublished PhD thesis, Carnegie Mellon University.

Wallman, K. K., and Harris-Kojetin, B. A. (2004), "Implementing the Confidential Information Protection and Statistical Efficiency Act of 2002," *Chance*, 17, 21–25.

Willenborg, L. C. R. J., and de Waal, T. (2001), *Elements of Statistical Disclosure Control*, New York: Springer–Verlag.

Yancey, W. E., Winkler, W. E., and Creecy, R. H. (2002), "Disclosure Risk Assessment in Perturbative Microdata Protection," in *Inference Control in Statistical Databases*, ed. J. Domingo-Ferrer, Berlin: Springer-Verlag, pp. 135–152.