NISS

Risk-Utility Paradigms for Statistical Disclosure Limitation: How to Think, But Not How to Act

Lawrence H. Cox, Alan F. Karr,
Satkartar K. Kinney

Technical Report 179
March 2011

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

Risk-Utility Paradigms for Statistical Disclosure Limitation: How to Think, But Not How to Act

Lawrence H. Cox, Alan F. Karr, Satkartar K. Kinney
National Institute of Statistical Sciences
Research Triangle Park, NC, USA

May 5, 2011

Abstract

Risk-utility formulations for problems of statistical disclosure limitation are now common. We argue that these approaches are powerful guides to official statistics agencies in regard to how to think about disclosure limitation problems, but that they fall short in essential ways from providing a sound basis for acting upon the problems. We illustrate this position in three specific contexts—transparency, tabular data and survey weights, with shorter consideration of two key emerging issues—longitudinal data and the use of administrative data to augment surveys.

Keywords: statistical disclosure limitation, risk-utility paradigm, disclosure risk, data utility, data quality, transparency, data swapping, tabular data, cell suppression, controlled rounding, survey weights, microaggregation

1 Introduction

Over the past fifteen years, risk-utility approaches to statistical disclosure limitation (SDL) have become pervasive. Our thesis here, as our title implies, is that risk-utility paradigms are useful for posing questions, but much less so for actually carrying out SDL. Put colloquially, risk-utility paradigms let us "talk the talk," but not "walk the walk." We support this thesis by means of several examples, each of which in itself raises important, unresolved issues in SDL, and which appear in §3–6.

An implication of our thesis is that risk-utility is not a scientific paradigm in the sense of Thomas Kuhn (Kuhn, 1962). In particular, as a discipline, SDL lacks fundamental characteristics of a science: a theoretical foundation subject to verification, falsification and generalization. Put bluntly, SDL is not (more optimistically, not yet) a science, but consists instead of a series of special cases connected only by common goals and common language.

2 Background

Here, we summarize risk-utility approaches to SDL. Essentially all of the content is contained in Figure 1, and the verbiage is largely an explication of it.

First, we introduce some additional terminology. The setting is data assembled by an official statistics agency charged with simultaneously protecting the privacy of the data subjects and the confidentiality of data values as well as making information available for government, research and other purposes. By *original data*, we mean the data as collected and processed by the agency. Processing may include edits that correct identifiable errors, imputation—perhaps multiple imputation—of missing values and weighting

class adjustments for nonresponse. Even so, substantial data quality problems may remain. By *masked data*, we mean the data made available by the agency following application of SDL. The degree of masking ranges from minimal—if the data are made available only to vetted users by means of restricted use data agreements—to substantial—if the data are released publicly.

Two further abstractions are useful: a *legitimate user* is one who employs the data for policy, research or other purposes with no intention of violating confidentiality or privacy. An *intruder*, by contrast, seeks to use the data in ways that violate privacy and confidentiality, most concretely by seeking to identify data subjects or learn the values of sensitive attributes.

The central tenet is that an official statistics agency releasing[1] information derived from confidential data faces a *decision problem*: it must make a tradeoff between the contradictory goals of decreasing disclosure risk and increasing data utility. That decision takes the form of choosing among multiple candidate versions of the released information. For concreteness, these may be thought of as arising from different choices of SDL methods, different choices of their associated parameters, and different modes of access to the data. To illustrate, for categorical data, given the choice of data swapping as the SDL mode, one candidate is associated with each choice of the attributes to be swapped, the swap rate, and any constraints imposed on swap partners. In Figure 1, these candidates correspond to the points in the scatterplot.

The next step is to assign to each candidate quantified values of *disclosure risk* and *data utility*,[2] which are plotted in Figure 1. Higher risk is generally, but not uniformly, associated with higher utility. What these measures are—Risk of what and to whom?, Utility in what sense and to whom?—are fundamental questions that lie at the core of our thesis. That "one person's risk is another person's utility" lies at the heart of the problem. There is also the pervasive problem that utility measures are either so broad that they become too blunt or so narrow that they fail to generalize (Woo et al., 2008, 2009).

Notwithstanding this difficulty, suppose the (risk,utility) values for the candidate releases are calculated, and plotted as in Figure 1. Then, a dramatic simplification of the decision problem results. Any candidate for which there is another candidate to the "southeast," meaning that the latter has both higher utility and lower risk, is automatically ruled out. The remaining candidates constitute the *risk-utility frontier*, and are connected in Figure 1 by the dotted lines. The frontier concept, which is analogous to efficient frontiers in economics, is one of the central contributions of the paradigm: regardless of the decision criteria employed by the agency, only candidates on the frontier need be considered.

Exactly how the agency does make its decision is at least agency- and dataset-specific. Even so, under any reasonable assumptions, curves of constant "value" to the agency have the convex shape shown in Figure 1, with value decreasing toward the northwest. Therefore, the optimal choice, shown by the large point in Figure 1, corresponds to the smallest value for which the curve intersects the frontier. One mathematical detail: convexity of the "iso-value" curves matters, but that of the frontier does not.

Given the utter simplicity and clarity of this paradigm, why is it so problematic? First of all, it fails to provide a scientific distinction between legitimate users and intruders. This is a serious problem because the soundest perspective might be that disclosure risk is identical to intruder utility. An operational distinction is proposed in §3.

The second problem is that the fundamental quantities of disclosure risk and data utility are neither precisely defined nor unambiguously measurable. In physics, basic characteristics such as mass, length and time are defined very precisely, even if measuring them may be challenging in some circumstances and

[1] In any form, from public microdata to access in a restricted data center.

[2] In this paper, data utility is synonymous with data quality.

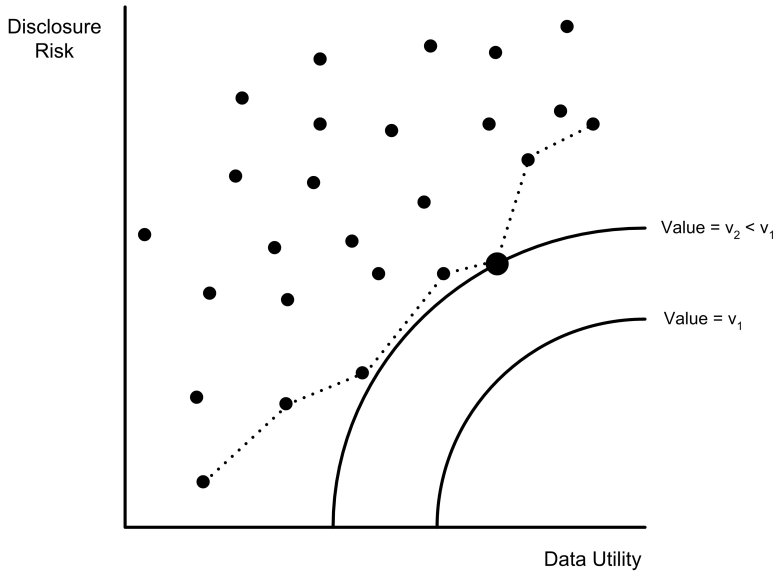

Figure 1: Pictorial representation of risk-utility paradigms. Points correspond to different candidate masked datasets. Those connected by dotted lines comprise the risk-utility frontier. The parallel, convex curves emanating from the lower right-hand corner represent constant value to the statistical agency. Given this value function, the large point is the optimal candidate.

error-free measurement may be impossible.

A third and related issue is the multiplicity of stakeholders, which minimally include the agency, the data subjects and users of the data. Neither of the latter is homogeneous. Some data subjects may be willing for their data to be published, while others will refuse to participate if they fear that their privacy may be compromised. Some users may wish only tabular summaries, graphs and maps, while others may want to perform sophisticated statistical analyses involving complex models, requiring variable transformations and producing detailed results and associated uncertainties.

A fourth and final issue is that risk in "disclosure risk" is used in an unconventional sense. Many quantifications of disclosure risk are (estimated) record-level probabilities of correct re-identification of a data subject from the masked data. There is no consideration of the harm associated with such re-identification.[3] Neither is there any attention to incorrect re-identification and its consequences. Equally, risk to the agency, to its employees and to data subjects are not at all the same thing (Trottini, 2001, 2003).

So, why not abandon the risk-utility paradigm entirely? The sections that follow are meant to be a dispassionate look at this question. Paraphrasing Eugene Wigner, we highlight both the "unreasonable effectiveness" of risk-utility paradigms and their apparent limitations.

[3] In the US and other countries, this results in part from the legal framework associated with official statistics, which defines disclosure to be an offense regardless of whether there is harm.

3 Transparency

This section articulates the role of *transparency*—that is, the release of information about processes and even parameters used to alter data—in SDL. In the process, a mathematical distinction between legitimate data users and intruders emerges, addressing one problem raised in §2.

3.1 What is Transparency?

Virtually without exception, statistical agencies have refused to divulge details concerning SDL procedures that have been applied in order to produce public microdata releases. For instance, if data swapping has been employed, an agency would not make public either which variable(s) had been swapped or the swap rate. The rationale reflects entirely a risk perspective: releasing such details is deemed to be risky. To date, there has been no quantification of the utility to legitimate data users if such information were to be released. As a result, risk-utility tradeoffs of the kind described in §2 are impossible.

Transparency is the extent to which a statistical agency releases information about the SDL processes used to transform the original, confidential database to the masked, released one. This idea arises in part from cryptography, where it is a fundamental precept (Kerckhoffs' principle) that encryption methods depend not on secrecy of the algorithm, but only on secrecy of keys. Indeed, when "breaking" the algorithm entails solution of a mathematical problem believed to be hard, such as factoring the product of two very large prime numbers, releasing the algorithm constitutes a deterrent to "computationally rational" intruders.

To illustrate, suppose that the original database O contains only numerical attributes and that the SDL consists of adding normally distributed noise, as in Karr et al. (2006a) and Oganian and Karr (2006). Let $\hat{\Sigma}$ be an estimator of the covariance matrix of O, such as the empirical estimator or a shrinkage estimator. Then each record $X_i \in O$ is replaced in the masked database M by

$$X_i^* = \frac{X_i + c\varepsilon_i}{\sqrt{1+c}}, \tag{1}$$

where the ε_i are independent with multivariate normal distribution $N(0, \hat{\Sigma})$ and c is a parameter selected by the agency. Typically, c is rather small: for example, 0.15 is employed in Oganian (2003).

Then, together with $M = \{X_i^*\}$, the agency may disclose, in order of increasing detail—and therefore also increasing risk *and* increasing utility:

1. Nothing.
2. That M was created from O by addition of mean zero noise.
3. That M was created from O by addition of normally distributed, mean zero noise.
4. That M was created from O by addition of normally distributed, mean zero noise whose covariance is $\hat{\Sigma}$.
5. That M was created from O by addition of normally distributed, mean zero noise whose covariance is $\hat{\Sigma}$ *and* the value of c.

By comparison with Alternative 2,

- Alternative 5 increases utility of the data dramatically. Legitimate users may perform principled inference for O using measurement error models.

- Alternative 4 is of value, although clearly less than Alternative 5. Because of the scaling in (1)—without which not releasing the value of c makes no sense—estimation of $\hat{\Sigma}$ from M is possible, which allows at least some analyses.
- Alternative 3 seems to offer no useful additional information beyond Alternative 2, which itself is not demonstrably more useful than Alternative 1.

So what is the agency to do? The major utility gain is between Alternatives 3 and 4. There is meaningful evidence (Karr et al., 2006a; Oganian and Karr, 2006) that addition of noise is less effective than other SDL methods—notably, microaggregation—at reducing risk. On the other hand, it is arguable that the four alternatives do not differ dramatically with respect to risk, since disclosing $\hat{\Sigma}$ releases only a highly aggregated characteristic of D. Therefore, invoking caution in the face of incomplete understanding, Alternative 4 might be chosen.

But, is this a scientifically-based decision, or simply one that is better than Alternative 1? The risk-utility paradigm showed the right question to ask, but provided only limited insight about how to answer it.

3.2 Distinguishing Legitimate Users and Intruders

Responding to one of the problems raised in §2, this setting permits us to formulate and illustrate an operational distinction between legitimate users and intruders: *legitimate users average, but intruders maximize*. We also introduce a strongly computational perspective on SDL, foreseeing a world in which sufficient computational power exists to consider all possible versions of O.

Let M denote the masked data, and let $\mathcal{K}$ be the knowledge about the SDL process *released by the agency*, which *at a minimum* includes the masked data M. We do not attempt to account for external knowledge on the part of either intruders or legitimate users. That is, $\mathcal{K}$ consists *only of knowledge released by the agency*.

We propose that both legitimate users and intruders wish to calculate the posterior distribution $P\{O = o|\mathcal{K}\}$, but *use this conditional distribution in fundamentally different ways*. Specifically, legitimate users wish to perform statistical analyses of the masked data M, as surrogates for analyses of O. Conditional on O, the results of such an analysis are a deterministic function $\mathbf{f}(O)$, which in general is vector-valued. To illustrate, for categorical data, $\mathbf{f}(O)$ may consist of the entire set of fitted values of the associated contingency table under a well-chosen log-linear model. In symbols, given $P\{O|\mathcal{K}\}$, legitimate users *integrate* to estimate $\mathbf{f}(O)$:

$$\widehat{\mathbf{f}(O)} = \int_{\mathcal{O}} \mathbf{f}(o)\, dP\{O = o|\mathcal{K}\}, \tag{2}$$

where $\mathcal{O}$ is the set of possible values of O. It is important to note that $\mathcal{O}$ depends on $\mathcal{K}$, even though the notation suppresses the dependence.

By contrast, intruders are not interested in integrals of the form (2), but rather in global or local maxima in $P\{O = \cdot|\mathcal{K}\}$, which correspond to high posterior likelihood estimates of the original data O. In the extreme, intruders would *maximize*, calculating

$$O^* = \arg\max_{o \in \mathcal{O}} P\{O = o|\mathcal{K}\}. \tag{3}$$

We do not prescribe what intruders would do using O^*, but assume only that this is whatever bad thing would be done using O itself, for instance, re-identifying records by means of linkage to an external database containing identifiers.

This distinction allows the agency to reason in principled manner about risk and utility, especially in terms of how they relate to $\mathcal{K}$:

High utility means that the integration in (2) can be performed or approximated relatively easily.

Low risk means that the maximization in (3) is difficult to perform or approximate.

As illustrated in §3.3, a central question is then: How large is the set $\mathcal{O}$ of possible values of O given $\mathcal{K}$? Of course, high utility and low risk remain competing objectives: when $\mathcal{O}$ is very large, then the maximization in (3) is hard, but so is the integration in (2). However, maximization typically becomes hard faster than the integration as $\mathcal{K}$ is decreased.

3.3 A More Detailed Example

In this section, we use doubly random data swapping (DRDS) to illustrate the formulation in §3.2. Assume the that original database O consists of n records, each containing p categorical attributes, and that the only form of SDL applied to the data is DRDS: both the records for which attributes are swapped *and* which attributes are actually swapped for each pair are randomized (Denogean et al., 2007).

Let $(p(j))_{j=1,\ldots,p}$ be the distribution of choice of the swap attribute in DRDS. Let M denote the masked data, and let $\mathcal{K}$ be the knowledge about the DRDS process *released by the agency*, which *always* includes the masked data M. Let k be the actual number of swapped *pairs* in M, which may or may not be part of $\mathcal{K}$. Mainly, we examine the effects of different choices of $\mathcal{K}$. In the hierarchy below, $\mathcal{K}$ becomes progressively larger, representing less transparency by the agency.

Case 0: $\mathcal{K}$ = Exact knowledge of which pairs of records and for each, which attributes were swapped. This extreme case is artificial because the effects of the swapping are exactly reversible. Mathematically, $\mathcal{O} = \{O\}$, and both (2) and (3) become trivial.

Case 1: $\mathcal{K}$ = Knowledge of which pairs of records were swapped and values of the $p(j)$, all of which are positive. For clarity, we examine this case in detail. By "which pairs of records were swapped" we mean that $\mathcal{K}$ contains a listing $(i_1, j_1), \ldots, (i_k, j_k)$ meaning that record i_1 was swapped with record j_1, ..., record i_k was swapped with record j_k. In effect, then, all that is not known is which attribute[4] was swapped in each pair. Because $p(j) > 0$ for all j, any attribute could have been swapped. Therefore, each swapped pair (i_ℓ, j_ℓ) has p possible antecedents in O, so that in the worst case,

$$|\mathcal{O}| = k^p. \tag{4}$$

(Recall that p is the number of attributes.) There are two reasons why (4) is an overestimate. First, not all k^p antecedents of a pair of records are distinct. For example if the records are

$$\begin{aligned} i &= \texttt{(Male, White, 20-25)} \\ j &= \texttt{(Female, White, 15-20),} \end{aligned}$$

then there are 2 rather than 3 possible antecedents:

$$\begin{aligned} i &= \texttt{(Female, White, 20-25)} \\ j &= \texttt{(Male, White, 15-20)} \end{aligned}$$

[4]For simplicity, we assume that only one attribute is swapped for each pair. Swapping of multiple attributes adds complication without adding insight.

and

$$\begin{aligned} i &= \texttt{(Male, White, 15-20)} \\ j &= \texttt{(Female, White, 20-25),} \end{aligned}$$

corresponding to having swapped "Sex" and "Age," respectively. This leads to a second version of (4):

$$|\mathcal{O}| = \prod_{\ell=1}^{k} \quad \text{Number of attributes on which } i_\ell \text{ and } j_\ell \text{ differ.} \tag{5}$$

Even this version, however, is not quite right: labeling of records in O is assumed to carry no information, so when two records differ in exactly two attributes, there is only one antecedent.

Returning to the main argument, since the $p(j)$ are included in $\mathcal{K}$, for each $o \in \mathcal{O}$

$$P\{O = o|\mathcal{K}\} = \prod_j p(j)^{\nu(j,M,o)}, \tag{6}$$

where $\nu(j, M, o)$ is the number of swaps (to create o from M) in which attribute j is swapped.

Where the bounds of the computational feasibility of calculating all elements of $\mathcal{O}$ lie in this case is not completely clear, but for realistic numbers such as $k = 100{,}000$ (say, in a data set of $n = 10{,}000{,}000$) and $p = 20$, $\mathcal{O}$ is already impossibly large relative to today's computational capabilities.

Case 2: $\mathcal{K}$ = Knowledge of which pairs of records were swapped and that $p(j) > 0$ for each j. What differs here from Case 1 is that (6) is replaced by

$$P\{O = o|\mathcal{K}\} = \frac{1}{k^p} = \frac{1}{|\mathcal{O}|}, \qquad o \in \mathcal{O}. \tag{7}$$

In (7) and below, $|S|$ denotes the cardinality of the set S.

Alternatively and more generally, legitimate users or intruders may place a Dirichlet prior distribution π on the $p(j)$, and replace (7) by a variant of (6):

$$P\{O = o|\mathcal{K}\} = \int \prod_j q(j)^{\nu(j,M,o)} d\pi(q). \tag{8}$$

By contrast, a "flat" prior π reduces (8) to (7).

Case 3: $\mathcal{K}$ = Knowledge of k and the values of the $p(j)$. Now, $\mathcal{O}$ becomes much larger than in Case 2:

$$|\mathcal{O}| = \left(\begin{array}{c} n \\ 2k \end{array} \right) \times \left(\begin{array}{c} 2k \\ k \end{array} \right) \times k!, \tag{9}$$

and presumably this case lies outside the realm of computational feasibility. For the same reasons that (4) overestimates $|\mathcal{O}|$, so does (9). However, (6) remains valid.

Case 4: $\mathcal{K}$ = Knowledge of k and that $p(j) > 0$ for all j. This case stands in the same relationship to Case 3 as Case 2 does to Case 1. The relevant formulas are (9) and

$$P\{O = o|\mathcal{K}\} = \frac{1}{|\mathcal{O}|}, \qquad o \in \mathcal{O}. \tag{10}$$

The corresponding variant of (8) is obvious.

Case 5: $\mathcal{K}$ = Knowledge of the values of the $p(j)$, but not of k. One must adjust (9) for the unknown value of k, which can range from 0 to $\lfloor n/2 \rfloor$:

$$|\mathcal{O}| = \sum_{\ell=0}^{\lfloor n/2 \rfloor} \left(\begin{array}{c} n \\ 2\ell \end{array} \right) \times \left(\begin{array}{c} 2\ell \\ \ell \end{array} \right) \times \ell!. \tag{11}$$

The associated variants of (6) and (8) remain valid.

Case 6: $\mathcal{K}$ = Knowledge that $p(j) > 0$ for all j. For this case, (11) remains true, so given no knowledge of m or the $p(j)$, (7) would apply. Given a joint prior on m and the $p(j)$, a variant of (8) holds.

It is worth examining the progression from Case 1 to Case 6. Two different things happen, although not simultaneously. First, $|\mathcal{O}|$ increases. This happens in Case 3 as compared to Cases 1–2 and in Case 5 as compared to Cases 1-4. Second, specific values of the $p(j)$ are withheld. This happens in Case 2 as compared to Case 1, Case 4 as compared to Case 3 and Case 6 as compared to Case 5. In these instances, a prior is placed on what is no longer known, and an earlier formula is integrated with respect to that prior. It may be that in terms of computation the increase in $|\mathcal{O}|$ is more burdensome.

An intriguing question is whether the techniques and software tools from algebraic statistics ((Pistone et al., 2001); see also §4) are relevant in this setting. For categorical data, DRDS can be formulated solely in terms of contingency tables. Because DRDS preserves only the one-dimensional marginals of tables, the largest choice of $\mathcal{O}$ is the set $\mathcal{O}_{\max}$ of all tables with the same one-dimensional marginals as M. However, it is not clear, except possibly for two-dimensional tables, whether all tables in $\mathcal{O}_{\max}$ are reachable from M by means of DRDS. See also §4.

3.4 Computational Issues

Allusion was made in §3.3 to the possibility that as $\mathcal{O}$ increases, the maximization in (3) becomes harder faster than the integration in (2). In other words, that it may be possible to thwart intruders computationally with less harm to legitimate users. Is this reasoning valid?

There is one compelling reason why the approach formulated here can work: it is possible to approximate the integral in (2) to verifiable accuracy by simulating from the distribution $P\{O = \cdot|\mathcal{K}\}$, but not equally possible to do the maximization in (3).

In particular, approximation of the integral in (2) can be done without computing $P\{O = o|\mathcal{K}\}$ for many more, or even all, values of o. On the other hand, absent an extremely clever algorithm and additional knowledge, approximation of the solution to (3) is not possible without allowing for the need to compute $P\{O = o|\mathcal{K}\}$ for all o. Whether existing techniques for simulation, such as those in Diaconis and Sturmfels (1998), are sufficiently powerful is a question to be investigated. There is also a sense in which (3) is inherently more challenging than (2), because even if all of the values of $P\{O = o|\mathcal{K}\}$ were known, there still is the need for an algorithm to perform the maximization.

Neither legitimate users nor intruders are defeated by storage requirements. If all elements of $\mathcal{O}$ can be generated sequentially, there is no need to retain values of $P\{O = o|\mathcal{K}\}$. The integration in (2) can be performed by computing the current $P\{O = o|\mathcal{K}\}$ and adding $f(o)P\{O = o|\mathcal{K}\}$ to a running sum, while the maximization in (3) can be done by comparing the current $P\{O = o|\mathcal{K}\}$ to the current maximum and storing o and $P\{O = o|\mathcal{K}\}$ only if $P\{O = o|\mathcal{K}\}$ exceeds the current maximum.

3.5 Other Approaches

The Bayesian perspective underlying §3.2 is not the only way legitimate users and intruders can make use of $\mathcal{K}$. Instead, it is possible to "Alter the analysis to accommodate $\mathcal{K}$." To make this clearer, consider Case 1 in §3.3. As an alternative to confronting the computational difficulties associated with (2), legitimate users might instead simply discard the records known to have been swapped and conduct analyses on the remaining $n - 2k$ data records. This approach seems especially appealing if $2k \ll n$ and if the user knows from $\mathcal{K}$ that all records are equally likely to have been chosen to be swapped. The only "price paid" is an increase in uncertainties. This strategy would work equally well when $\mathcal{K}$ is smaller than in Case 1. All it requires is to know which records were swapped, and not either how the swapped records were paired or the values of the $p(j)$.

More generally, legitimate users might attempt to estimate from M and additional information in $\mathcal{K}$ the probability that each record in M had been swapped and to weight the records for analysis purposes inversely to these probabilities. The analysis procedure in that case would be similar to those designed to handle sampling weights. The "throw out known swapped records" strategy is simply an extreme version of this one, because it in effect assigns weight zero to discarded records.

But, there are still other ways to use estimated probabilities that records were swapped. In particular, such estimates might be used to construct prior distributions in §3.3. To make this clearer, consider Case 2, where a prior on the values of the $p(j)$ is needed in (8). Some implementations of DRDS use attribute selection probabilities $p(\cdot)$ that reflect the extent to which each attribute is independent of the others. A legitimate user or intruder who knew this information from $\mathcal{K}$ might estimate the probabilities in the same way from M, and use the posterior distribution for the $p(\cdot)$ as the prior distribution π in (8).

3.6 Transparency: Lessons Learned

The principal contribution of the risk-utility perspective with respect to transparency is to symmetrize the role of the knowledge $\mathcal{K}$ with respect to risk and utility: information in $\mathcal{K}$ about the SDL has *both risk and utility consequences*. The risk-utility perspective also makes the resonance of transparency with cryptography more compelling. Both of these are rather clearly "how to think."

But, the absence of insight about "how to act" is equally striking. The focus on $P\{O = \cdot|\mathcal{K}\}$ as the object of interest to both legitimate users and intruders is intriguing, but without specifics does little more than substitute mathematical symbols for words. This gap is more glaring when consideration of external information in $\mathcal{K}$ is necessary.

Likewise, the assertion that "legitimate users integrate, intruders maximize" remains unsubstantiated, and may not be a scientific statement in that sense that it can be tested by means of experiment, and it is probably an oversimplification.

Finally, whether, as §3.3 intimates, computational complexity is a path to measuring risk and utility is a question that cannot be answered without going outside risk-utility paradigms.

The extent to which the same lessons learned recur in subsequent sections is striking. In §4, the issue of "all possible values of O" is a central, amplified need to assess the extent to which elements of $\mathcal{O}$ differ from one another.

4 Tabular Data

In this section, we examine issues of SDL for tabular data.

Tabular data arise as follows. A partially ordered collection of *tabular cells* $\{X_j\}_{j \in J}$ is defined. Each respondent is assigned to a unique minimal member (cell) of the collection. The cell value is defined to be the sum over respondents in the cell of a particular data value pertaining to the respondent. For purposes here, respondent data are nonnegative count or magnitude data. Tabular data may arise as familiar one-way, two-way and three- or higher-way (dimensional) tables, hierarchies or partial orders over sets of tables, and linked tables. Statistical agencies have been releasing tabular data as hard copy for decades, and now via on-line query systems.

We adopt the following notation. For each unit i in the population or sample, $1 \leq i \leq I$, and each cell j in the tabular system, $\alpha(i, j) = 1$ if unit i contributes to cell j, and $\alpha(i, j) = 0$ otherwise. If $\alpha(i, j) = 1$, the contribution of unit i to cell j is $x_{ij} \geq 0$; if $\alpha(i, j) = 0$, $x_{ij} = 0$. For count data, $x_{ij} = \alpha_{ij}$. The *cell value* is $x_j = \sum_{i:\alpha(i,j)=1} x_{ij}$. For notational convenience and without loss of generality, we assume contributions in cells are in nonincreasing order: $x_{i(j),j} \geq x_{i'(j),j}$ whenever $i'(j) \geq i(j)$. Disclosure risk is based upon modeling intruder behavior and establishing criteria and procedures to thwart unsafe intrusion. Criteria for successful SDL for tabular data include:

- In addition to the original tabulations, that there exist a sufficient number of *alternative tabulations* that could have resulted in the same *masked tabulations*. This is precisely the same reasoning that underlies §3.3.
- The posterior predictive probability of the original tabulations conditional on the masked tabulations is small. See §3.2.
- Among these alternative tabulations, sufficiently many alternative values for each sensitive cell value are exhibited; alternative values sufficiently distant from each sensitive cell value are exhibited; and the posterior predictive probability of a sensitive value conditional on the masked tabulations is small.

Some of these overlap with concepts for microdata such as k-anonymity and 1-diversity. Recent research has raised challenges or gaps to concepts and theory underlying tabular SDL and to the quality of SDL-treated tabular data, as well as related transparency issues. This section focuses on these issues. In particular, while in many ways natural, these criteria can be difficult to assure or even verify.

4.1 A Theory for Risk and Disclosure Limitation in Tabular Data

Once the statistical agency has established criteria for identifying disclosure and assessing the adequacy of SDL procedures, it expresses them in quantitative form as disclosure rules: "Unsafe disclosure occurs if . . ., and is absent otherwise." Next the agency identifies suitable SDL methods, and finally applies the SDL to original tabulations with sufficient intensity to assure that the masked data are disclosure-free. Modern SDL methods incorporate features aimed at preserving the quality of original tabulations in the masked tabulations.

A theory for tabular SDL based upon limiting risk of disclosing individual respondent data through narrow (unsafe) estimates of corresponding (sensitive) cell values was developed and implemented in the late 1970s, and soon followed by several papers dealing with specific tabular SDL methods (Cox, 1980, 1981, 1987a, 1995; Cox and Ernst, 1982; Kelly et al., 1992). During the 2000s, theory was extended to

measure and control effects of SDL on data quality and usability (Cox, 2008; Cox and Kim, 2006; Cox et al., 2004, 2006), increasingly computationally efficient approaches to potentially intractable problems (Kelly et al., 1992; Cox, 2003) were introduced (Cox, 1995; Fischetti and Salazar-Gonzalez, 2001). An earlier methodology for tabular SDL addressing only exact disclosure of sensitive cell values was provided by Fellegi (1972).

The underlying theory (Cox, 1981) is based on quantitative methods to identify and measure disclosure, which in turn provide a lower bound on distortion necessary to achieve successful disclosure treatment, as follows. A *linear sensitivity measure* S is a linear functional $S(X_j) = \sum_{i=1}^{I} w_i x_{ij}$, normalized so that $w_I = -1$. Cell X_j is *sensitive* if and only if $S(X_j) > 0$. Familiar sensitivity measures are:

For count data, the t-threshold rule: $S\left(X_j\right) = (t-1)x_{1j} - \sum_{i=2}^{I} x_{ij}$.

For magnitude data, the p-percent rule: $S_{p\%}(X_j) = (p/100)x_{1j} - \sum_{i=3}^{I} x_{ij}$; and the p/q-ambiguity rule: $S_{p/q}(X_j) = (p/q)x_{1j} - \sum_{i=3}^{I} x_{ij}$.

These rules express established notions of disclosure risk for tabular data: for count data, small counts allow the intruder to pinpoint or infer with confidence the traits of a respondent; and, for magnitude data, competitors will seek to estimate another's data to within a small percentage, often with the aid of prior information, such as their own contribution to the sum.

The theory of sensitivity measures assures two important conditions: (1) if X_j is sensitive, then the minimum value of a nonsensitive cell $X_{j'}$ containing X_j equals $x_{j'} = x_j + S(X_j)$, and (2) if $w_i \geq w_{i'}$ for $i' > i$, then for $X_{j'} = X_j \cup X_{j''}$ we have $S(X_{j'}) \leq S(X_j) + S(X_{j''})$. The first condition provides an operational minimum for the amount of disclosure limitation needed to treat X_j; the second, known as *subadditivity*, assures that the union of two nonsensitive cells is also nonsensitive, which is important for both operational and conceptual reasons.

The theory of linear sensitivity measures has driven disclosure limitation methodology and practice in tabular data for several decades. For rounding, it assures that all cells rounded to base t or greater will be nonsensitive. For perturbation at the cell or microdata level, it provides the framework to which the probability distribution for random errors is determined. For complementary cell suppression and controlled tabular adjustment, it provides the operational parameters (protection limits) that drive binary SDL decision-making algorithms.

4.2 Effects on Data Quality and Usability

The principal methods for tabular SDL are: cell value rounding; cell value or microdata perturbation; complementary cell suppression; and controlled tabular adjustment. Quality is assessed based on preserving additivity, accuracy, consistency and usability. Consistency of cell values refers to ensuring that equivalent true cell values are masked equivalently; consistency of inference refers to ensuring that statistical analysis based on true and masked values result in equivalent inferences. Usability is assessed by preserving in masked values the range and ease-of-use of analytical methods applied to true values.

Rounding. Statistical data may be rounded for various reasons. For count data, replacing all cell values by multiples of the threshold t (rounding base t) has the effect of reducing the accuracy of intruder inferences of true values to within a range of t units. This prevents the t-threshold rule from achieving positive values and thus, by definition, limits disclosure risk to an acceptable level. Only rounded values are released.

Conventional rounding is optimal with respect to accuracy. Unfortunately, conventional rounding does not preserve additivity: $3 + 4 = 7$ but, rounding to base 5, $5 + 5 \neq 5$. The simplest relaxation of conventional rounding is adjacent rounding: round each unrounded value to either of its two adjacent multiples of the base. Controlled rounding (CR) is rounding that preserves additivity. For one- and two-way tables, adjacent rounding can be controlled (but in general not for higher-way tables), and rounded values (zeroes, in particular) can be preserved (Cox and Ernst, 1982). Thus, $3 + 4 = 7$, and $5 + 5 = 10$, or $0 + 5 = 5$, or $5 + 0 = 5$, but neither $0 + 0 = 5$ or 10 nor $0 + 5$ or $5 + 0 = 10$. Statistically unbiased controlled rounding is also possible (Cox, 1987b). Rounding can be ineffective for disclosure limitation purposes in magnitude data due to typically skewed distribution of cell values, but in some cases may be applied selectively. In terms of data protection, properly implemented controlled rounding provides ideal disclosure limitation—a uniform predictive posterior distribution for true values given rounded values (Cox and Kim, 2006). In terms of data quality, controlled rounding preserves additivity, accuracy, and usability well. Open question are how to assure consistency under controlled rounding and how to relax rounding constraints to assure additivity when controlled rounding fails.

Random Perturbation is based on introducing random noise into cell values. For count data, the noise is integer, typically small, with zero mean, and is added to the true cell value x_j, resulting in the masked value. For magnitude data, noise is multiplicative, with unit mean, and applied to the underlying microdata x_{ij}, resulting in the masked value (Evans et al., 1998). Random perturbation preserves expected values of cell values and totals, but for magnitude data can result in large deviations between detail and total cells. For establishment data, care needs to be taken to ensure that data are masked at both the establishment and enterprise level (often this is accomplished by ensuring that adjustments to individual establishments within an enterprise are all in the same direction: up or down). Perturbation is vulnerable to attack via repeated queries, unless steps are taken to keep responses to equivalent queries constant (Fraser and Wooton, 2005).

Complementary Cell Suppression (CCS) deletes (suppresses) from the tabulations any cell value failing the sensitivity measure. To thwart narrow estimation of suppressed values, additional, nonsensitive cell values must also be suppressed. Properly specified mixed integer linear programming (MILP) models for complementary cell suppression assure that full protection is achieved (Cox, 1980, 1995; Fischetti and Salazar-Gonzalez, 2001). Data quality is preserved through use of an appropriate information loss criterion (e.g., total value suppressed, total number of suppressed cells, Berg entropy) as the MILP objective. Computing an optimal or near-optimal suppression pattern is a complex (NP-hard) computational problem (Kelly et al., 1992) involving one binary integer decision variable (suppress or not) for each nonsensitive cell. Suppression degrades data usability for unsophisticated users due to missing cells, and also for sophisticated users as the suppressions, selected not on the basis of a proper missing data probability model, are difficult to reconstruct reliably (Cox, 2008).

Controlled Tabular Adjustment (CTA) was developed as an alternative to complementary cell suppression (Cox et al. 2004). Instead of suppressing sensitive cells, the true cell value x_j is replaced by either of its two adjacent safe values, viz., $x_j + S(X_j)$ or (typically) $x_j - S(X_j)$. Appropriate nonsensitive cell values are adjusted to restore additivity. The corresponding mixed integer-linear program has one binary decision variable for each sensitive cell—typically far fewer than for cell suppression (but still NP-hard). Accuracy can be assured by imposition of capacity constraints. Broader data quality can be assured analytically Cox et al. (2004) or in distributional terms (Cox and Kim, 2006).

$$\left(x_{+\mathrm{jk}}\right) = \begin{pmatrix} 2 & 2 & 2 \\ 2 & 1 & 2 \\ 2 & 1 & 2 \\ 2 & 1 & 2 \end{pmatrix}, \left(x_{\mathrm{i+k}}\right) = \begin{pmatrix} 2 & 2 & 0 \\ 1 & 1 & 0 \\ 2 & 0 & 2 \\ 3 & 0 & 1 \\ 0 & 2 & 2 \\ 0 & 1 & 3 \end{pmatrix}, \left(x_{\mathrm{ij+}}\right) = \begin{pmatrix} 2 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 2 & 2 & 0 & 0 \\ 0 & 0 & 2 & 2 \\ 2 & 0 & 2 & 0 \\ 0 & 2 & 0 & 2 \end{pmatrix}.$$

Table 1: $6 \times 4 \times 3$ table failing the interval property (De Loera and Onn, 2004, p. 344).

4.3 Challenges and Gaps

These challenges and gaps are united by the failure of risk-utility considerations to tell an agency what to do.

Sufficiently Many Alternative Values. The t-threshold rule is intended to mask small counts (e.g., 1 or 1 and 2) within a larger set of counts (e.g., $t = 2$ or $t = 3$, 4, or 5). Intuitively, if not precisely, the intruder can guess a true count (e.g., 1) with probability at most $1/t$. All that was left for the statistical agency to do was to choose a value for parameter t that best expressed its confidentiality policies and practices. All of this seemed quite sensible until De Loera and Onn (2004) discovered a new mathematical fact, presented here for tables.

The additivity constraints imposed by marginal totals and additional constraints related to the SDL method (e.g., adjacency conditions in CR; unsuppressed cell values in CCS), define a system of linear equality and inequality constraints over the masked values. To each masked value x_j, there corresponds a minimum integer value m_j and a maximum integer value M_j it can achieve subject to this system. It is well known that every continuous value in the interval $[m_j, M_j]$ is achieved in some feasible continuous solution of the system. De Loera and Onn (2004) showed that the corresponding property for integer solutions the interval property can fail, even for moderately sized 3-way tables. See Table 1: x_{111} can be 0 or 2, but not 1.

Sufficiently Many Alternative Tables, and Their Distribution. A valid risk criterion would be to assure that the probability of correctly guessing a masked value is small, meaning that the relative frequency of the true value among all feasible (integer) values is small. In addition to existence of sufficiently many alternative values and tables, the probability distribution for masked values should not be spiked at the true value. This is difficult to achieve in practice, for several reasons. First, counting tables and exhibiting alternative tables can be hard. Second, integer feasibility is not assured—there may be no tables satisfying certain conditions. And finally, cell and table probabilities are difficult to compute.

Other techniques from algebraic statistics (Pistone et al., 2001) may prove useful in addressing these issues. In particular, methods for computing integer moves from the true table to alternative tables would answer questions surrounding existence of sufficiently many alternative values, sufficiently many alternative tables and, combined with statistical methods, distributions of alternative values and tables.

Important SDL questions arise from developments presented in the two preceding subsections. First, the t-threshold rule fails to define disclosure in a manner that is consistent across tables and that assures predictable limits on disclosure risk. How might it be replaced or exceptions identified and handled? Second, the interval property is preserved for tables based on a decomposable graphical model (Dobra, 2003) and for tables of network type (Cox, 2007). In these cases the t-threshold rule remains sufficient. Should these tables be favored over those potentially failing the interval property? Finally, is it possible to compute a

minimal perturbation of a table failing the interval property, to achieve a table fulfilling the property, bearing in mind that minimal adjustments also can result in infeasibility (Cox, 2003)?

Methods for Analysis of Magnitude Data. Unlike count data, for which an extensive theory of log-linear models is available, methods for analysis of tabular magnitude data are not well established, making it difficult to measure utility and effectively balance quality and confidentiality.

Query Systems. Statistical database query systems potentially enable release of high numbers of high-dimensional tables or tabulation cells (Dobra et al., 2002), but they require SDL methods that are dynamic and have a long memory. Disclosure risk is no longer static but increasing with time. Techniques are needed for organizing past and current queries and efficiently evaluating current disclosure risk, assessing and controlling entropy creep, and assessing disclosure risk associated with omnibus policies such as answering only five- or lower-dimensional queries. Strong defenses to identify and thwart tracker attacks (Schlörer, 1980) are also required.

Optimality and Computational Obstacles. Many tabular SDL methods are based on mathematical optimization models. In some cases (CTA), the objective function provides computational convenience unrelated to data quality. In other cases (Berg entropy for CCS), the connection to quality of inferences and resultant decisions is opaque. Typically, there is no optimal solution and a great deal of computational energy devoted to solving to optimality NP-hard problems could be saved if criteria were in place to determine: when is a solution good enough?

Transparency. Rounding is completely transparent: all computations are in multiples of the rounding base, so no small values can be inferred. No parameters or procedures are hidden from the user. For this reason, the quality properties of rounding can be studied (Cox and Kim, 2006). Random perturbation can be weakened by transparency and repeated queries. An open transparency issue is whether to release parameters of the noise distribution to improve variance estimation or whether this release degrades the effectiveness and security of the SDL. If suppression is performed repeatedly—as in periodic surveys and censuses—or transparently, as by releasing parameters of the sensitivity rule– it can be vulnerable to intruder attack to the point of being undone (Cox, 2009). Controlled tabular adjustment can be performed transparently without risk.

4.4 Tabular Data: Lessons Learned

Some risk-utility paradigms are built upon a minimax notion: subject to assuring minimal confidentiality protection, data utility should be maximized. There are two resultant shortcomings.

First, the assumption that both disclosure risk and data quality can be measured and commensurably. Indeed, even if possible, measurement would be with error and consequently a solution on the frontier is likely to be unstable, arguing for finding good interior solutions. A more actionable framework than R-U would be to develop procedures that ensure safe releases of acceptable quality that lie in the interior of the safe region (Cox et al., 2004; Cox and Kim, 2006). If it is desirable and possible to improve such solutions, then this could be done up to an appropriate level of effort, bearing in mind the question: how good is good enough? This it seems is the key question to answer, as opposed to searching for optimal solutions.

Second, data quality is multi-dimensional (Karr et al., 2006b), and often the criteria compete. Just considering the quality dimension "accuracy," it is clear that for some data values SDL must blunt accuracy. For some users, individual values or cases are of primary concern. Consequently, an SDL method that alters the fewest number of original values might be preferred. Or, if the user is analyzing sets of values, such as

tables, or computing correlations between two variables, then methods preserving other quantities, such as totals or correlations, may be preferred. Attempting to develop a manageable number of objective functions or a multi-criteria methodology in this setting may be unrealistic.

Specifically for tabular SDL, in order to apply R-U effectively, we would need a quantitative framework for answering the following questions:

- Given two safe alternative tables for an original (unsafe) table, which one is safer and by how much? Hellinger distances (Gomatam et al., 2005) provide one—but excessively blunt—approach.
- Given two safe alternative tables for an original (unsafe) table, which one is closer in quality to the original, and by how much?

There are formidable theoretical and computational obstacles to answering these and many similar questions.

5 Weights

Survey weights are well-understood from a qualitative perspective in terms of utility, but have received much less consideration in quantitative terms of either utility or risk.

Nearly universally, surveys conducted by statistical agencies involve multi-stage sampling designs, and may also include over-sampling of particular subpopulations. Weights are an integral component of the dataset. They typically begin as inverses of the probability of selection, but are often adjusted in order to reduce nonresponse bias, as well as account for disruptions such as budget modifications and exogenous events and known differences between the sample and the frame.

Both unweighted and weighted analyses of the data are performed, the latter often targeted at construction of "national estimates," for example of the prevalence of a disease or the number of children achieving a particular level of educational proficiency. However, when data are altered for the purpose of SDL, the dataset may no longer accurately represent the population.

5.1 Risk-Utility Perspective

Existing research in SDL has done relatively little to address how weights can be included safely in public-use microdata (Fienberg, 2009). There is consensus that one principal risk from inclusion of weights in microdata releases is that weights can be used to deduce information about design variables that are not released (De Waal and Willenborg, 1997; Willenborg and de Waal, 1996).

To illustrate, if weights are a function only of strata, then records with the same weight come from the same stratum, and it might then be easy to determine which stratum is which. Other risks include using weights to link records to primary sampling units (PSUs) even when geography is reported only very coarsely, and using replicate weights provided to facilitate variance estimation to re-identify suppressed design variables (Lu and Sitter, 2008; Park et al., 2006). If stratum-defining variables have been retained but altered for purposes of SDL—for instance, by swapping—weights may be used to detect or even reverse the SDL.

The values of weights can also be problematic. Very low weights are indicative of sample or even population uniqueness. At the other extreme, some health and education surveys contain self-representing PSUs, which are included with probability one, but in which sampling rates are small, and hence weights

are high.[5] In the system for geographic aggregation of data on chemical use by farms described in Karr et al. (2001), the weights were deemed by the statistical agency owning the data as more sensitive than the chemical use levels. In establishment data, weights could divulge proprietary data to competitors.

Attention to utility issues for weights is equally scant. The proposals in Willenborg and de Waal (1996) to employ subsampling and noise addition in order to reduce risk seem to engender an unacceptably large loss of utility. A few papers have proposed solutions addressing the utility of weights in specific situations. An example is Mitra and Reiter (2006), in which improved utility of weighted analyses is achieved in the context of partially synthetic data by recalculating sampling weights using synthetic design variables; however, risk issues are not addressed.

The adjustment of weights to attenuate known differences between the sample and the frame is often effected by means by population controls and poststratification adjustments. In the simplest case, weights are linearly re-scaled so that certain totals match information published elsewhere. However, matching the sample data to known population values can result in design variables that can be re-identified (Willenborg and de Waal, 1996, 2001). Fienberg (2009) takes the position that population controls be abandoned, in part because poststratification of weights to match them imposes unnecessary risk. Fienberg (2009) argues further for model-based rather than design-based analyses (see also Little (2003)), which would allow weights to be suppressed and obviate risks associated with them. However, release of design variables is itself risky, and in effect this proposal seems simply to replace one risk by another.

Obviously, unanswered questions abound. Should SDL methods be applied to design variables, survey responses, or both? Should they be applied to weights? What utility arises when unaltered weights are released with altered data? Can population controls safely be used? If weights are central to SDL, is the current absence of SDL considerations at the design stage injurious?

Once more, the risk-utility perspective is of evident value in posing questions, but of little (to date) demonstrable value in resolving them.

5.2 Indexed Microaggregation: Altering Weights for SDL

Following the model in §3.3, we now explore an example. The underlying rationale flows from §5.1: not modifying weights at all may increase disclosure risk—perhaps significantly, but modifying them in the wrong way may decrease utility—perhaps dramatically.

Many SDL techniques are inappropriate for weights. For instance, swapping weights can create records that can be recognized as having been swapped, adding to risk, at the same time that the utility of weighted analyses, as measured by the fidelity of analyses performed on the masked data M as compared to the same analyses performed on the original data O, can be decreased dramatically.[6] No matter how one interprets risk and utility, any process that both increases risk and decreases utility is bad. Similarly, if weights are highly discrete, addition of noise will decrease utility with no compensating impact on risk.

Instead, we consider a novel variant of multivariate microaggregation (Domingo-Ferrer and Mateo-Sanz, 2002; Domingo-Ferrer et al., 2006), which we term *indexed microaggregation*. Briefly, microaggregation is a SDL technique for numerical data in which data points are grouped into disjoint subsets of modest size

[5] Large metropolitan areas are typical examples of self-representing PSUs with low sampling rates.

[6] Consider what happens to estimates of national income when disparate weights are swapped between records with disparate income values. To address this issue, the Data Swapping Toolkit developed by NISS (National Institute of Statistical Sciences, 2003), which does accommodate weights, allows constraints on the disparity of weights associated with swap pairs. But of course, constraints that are too tight fail to reduce risk.

n (for example, $n = 3$), using a subset of the variables, and then within each group, the values of each of those variables are replaced by their within-group mean.[7] Microaggregation is effective at reducing risk, but attenuates variability in the data, which can be restored by addition of noise (Oganian and Karr, 2006).

The novelty is that we perform the grouping using response variables y, but average the associated weights w. The heuristic justification is that weighted analyses involving those variables will not be distorted even if weights are altered rather substantially.

Understanding whether this reasoning is valid currently seems approachable by means of simulation experiments. For these, we employed a database of 11,441 records in a public version of the National Health and Nutrition Examination Survey (NHANES), which is based on a cluster sampling design. The variables include INCOME, GENDER, AGE, STRATUM, PSU, EDUCATIONAL ATTAINMENT, and POVERTY-TO-INCOME RATIO (PIR). PIR was top-coded in the NHANES data, so we added random positive values to the top-coded values to simulate the confidential variable INCOME. In order to emulate the disclosure risk strategies in Willenborg and de Waal (1996), weights were assigned distinct and discrete values, shown in Figure 2, such that each unit in a stratum has the same weight.

Simulations were run with multiple protection strategies to protect the values of INCOME: top-coding, rank swapping, and microaggregation. The R package `sdcMicro` (Templ, 2008) was used to perform rank swapping, and standard microaggregation routines were employed. Means and standard errors were computed in each simulation using the R package `survey`. The altered weights w^* were constructed by ordering the variable INCOME and using its values to group the original weights into set of three. For example, the weights corresponding to the three lowest values of INCOME were replaced with their group mean.

We now discuss the results of the simulation experiment. In the original NHANES data, there are 28 strata ranging in size from 276 to 676. There are 28 distinct values that $w_i, i = 1, \ldots, n$ can have, one per stratum, ranging from 1 to 5000, which are shown in Figure 2. By contrast, in the masked data, stratum is masked or suppressed, and the $w_i^*, i = 1, \ldots, n$ take on 646 distinct values between 1 and 5000, with groups of units with the same value of w_i^* ranging from 1 to 129. Within each of the 28 strata, there is a great deal of variability in the values of w_i^*, as shown in Figure 3. As anticipated, this complicates re-identification of stratum from weight. However, it does not entirely remove the association between stratum and weight.

One way of evaluating risk is to estimate $P(\text{stratum}|\text{altered weight})$, which quantifies how difficult it is to identify the stratum from the altered weights. The "flatter" this distribution, the more difficult it is to re-identify, which is exactly the same reasoning appearing in §3 and §4. Figure 4 illustrates these probabilities for altered weights $w_i^* \in \{1000 \pm 100, 2000 \pm 100, 3000 \pm 100, 4000 \pm 100\}$. These results suggest that the risk of re-identification using sample sizes and known population values is low; however, there may be cases where outliers would still be at risk. It is also clear in Figure 4 that the strata fall into two classes having substantially differing weights, which may pose yet another risk.

To evaluate the utility of masked data M containing the altered weights, we compare a set of weighted analyses using both the original weights and the altered weights. For simplicity, we focus on weighted means and sums, since it is not always clear how to accommodate weights in more complicated estimands. Indeed, for regression models, a better approach is to incorporate variables affecting sampling design and nonresponse, if available, into the model rather than using weights (Fienberg, 2009; Gelman, 2007).

Using the unaltered INCOME and unaltered weights, INCOME has a mean of 2.6921 and a standard error of 0.0601. We compare this to the values obtained using the altered weights and unaltered INCOME, and

[7] Put differently, the multiple-variable values are replaced by group centroids.

	Unaltered Weights		Altered Weights	
Income	Mean	SE	Mean	SE
Unaltered	2.6921	0.0601	2.6929	0.0475
Top-Coded	2.5024	0.0459	2.5023	0.0381
Rank Swapped	2.6774	0.0615	2.6745	0.0491
Microaggregated	2.6907	0.0601	2.6913	0.0475

Table 2: Univariate response and discrete weights: Mean and Standard Error.

	Educ=1		Educ=2		Educ=3	
Weight	Mean	SE	Mean	SE	Mean	SE
Unaltered	1.7373	0.0529	2.5963	0.0461	3.6122	0.0891
Altered	1.7353	0.0454	2.5596	0.041	3.6226	0.0624

Table 3: Univariate response and discrete weights: INCOME by EDUCATION LEVEL.

repeat the comparison using different versions of altered INCOME obtained using top-coding, rank swapping, and microaggregation. These results, shown in Table 2, show that the mean of INCOME is well-preserved in all cases, although variance is underestimated. As noted above, this is not surprising: microaggregation reduces variability.

Although only INCOME was included in the strategy, analyses based on other variables can still be examined. For example, EDUCATIONAL ATTAINMENT is highly correlated with INCOME, and weighted analyses involving the two are well preserved. Table 3 shows the mean of INCOME disaggregated by EDUCATIONAL ATTAINMENT. Changes in means are minimal; as before, a reduction in variance is noted. Table 4 shows that the weighted distribution of education level is distorted slightly more, although the percentage differences are very small.

5.3 Weights: Lessons Learned

The simulation results in §5.2 illustrate that the indexed microaggregation can be useful when weights are deemed to be risky. Arguably, the strategy arose from a risk-utility perspective, which did in a sense suggest what to do. In addition, risk and utility served as a means of evaluating what was done. But, there is no evidence of generalizability, nor is there any way of knowing whether the risk and utility measures are the

Weights		Educ=1	Educ=2	Educ=3	Total
Unaltered	n	4592231	3027135	5079930	12699296
	%	36.16	23.84	40.00	100
Altered	n	4517756	3090182	5092194	12700132
	%	35.57	24.33	40.09	100

Table 4: Tabulation of INCOME by EDUCATIONAL ATTAINMENT.

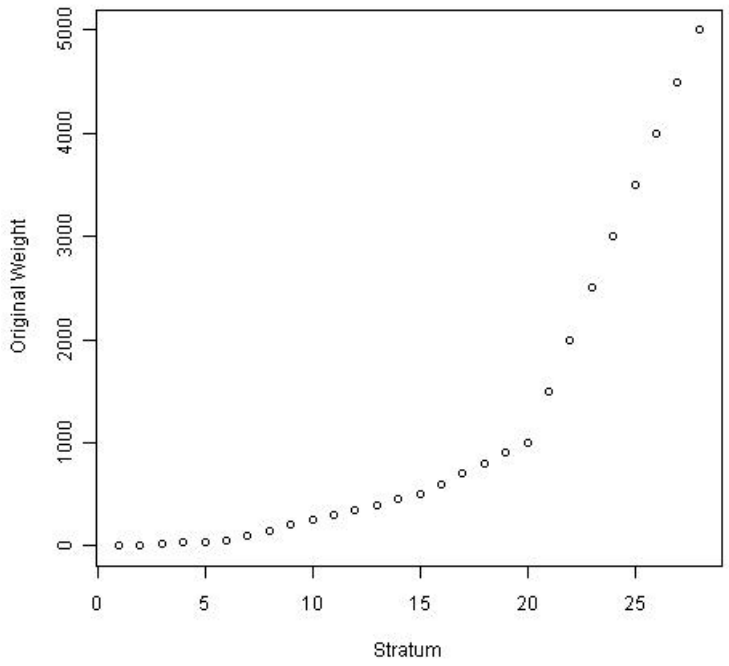

Figure 2: Plot of original weights by stratum.

"right ones." Moreover, not everything was done that needs to be done: in particular, the variability reduction caused by microaggregation was not addressed, a problem for which approaches exist (Domingo-Ferrer and í González-Nicolás, 2010; Oganian and Karr, 2006). So, just as in §3, the risk-utility paradigm helps pose questions, but those questions lie—perhaps permanently—outside its capability to resolve.

Among those are the questions at the end of §5.1, as well as a host of others. One of these is that many datasets have not one but several sets of weights. The Early Childhood Longitudinal Study–Kindergarten Class of 1998–99 (ECLS-K) contains dozens of sets of cross-sectional and longitudinal weights, reflecting such factors as additional data collections and attrition of data subjects. Indeed, many surveys have both sets of weights for longitudinal analyses and other sets for cross-sectional analyses.

6 Emerging Issues

In this section we discuss briefly two "emerging issues," although in other senses neither is new. What *is* changing is the urgency with which they need to be addressed, which highlights our current inability to do so.

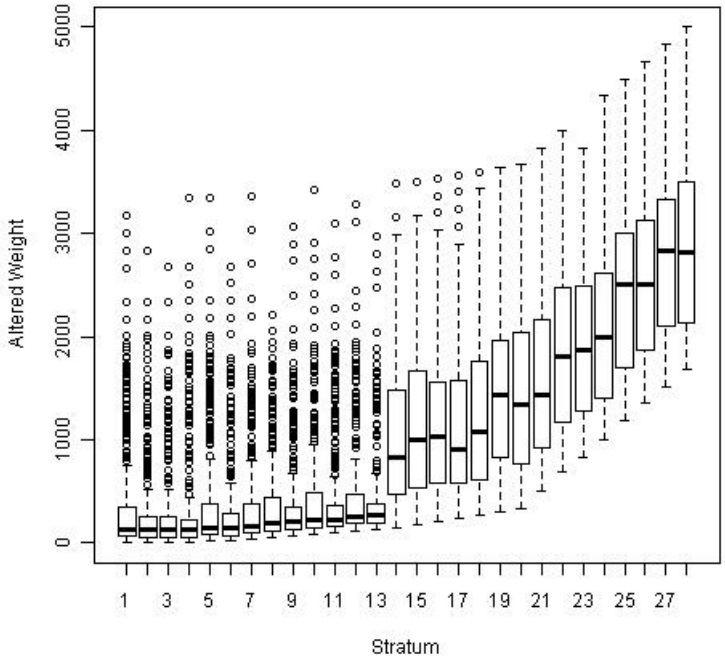

Figure 3: Plot of altered weights by stratum.

6.1 Longitudinal Data

Longitudinal data present an additional challenge to creating safe public-release files that preserve the longitudinal information. The longitudinal links and patterns provide additional information about individuals units that can be exploited by intruders. Agencies and researchers have long been aware of this risk (Cox and Zayatz, 1995; National Research Council Committee on National Statistics, 2000; Zayatz et al., 1999). However, risk formulations typically assume a cross-sectional context and do not address longitudinal linkages. The time series structure of the data can also be used in a Bayesian formulation to compromise SDL (Holan et al., 2010).

There is little guidance on SDL for longitudinal data. The typical path taken by agencies releasing longitudinal data is to impose access restrictions, so that unrestricted longitudinal datasets are rare. Recent work using synthetic data methods has yielded some results, though without formal risk–utility evaluations. For example, Abowd and Woodcock (2001, 2004) describe the problem of protecting longitudinal data that cross multiple sampling frames, such as employers and workers in the US Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) program.

Kinney et al. (2010) generated a *synthetic data* version of the US Census Bureau's Longitudinal Business Database (LBD). In this context, "synthetic data" refers to replacement of a portion (Typically, some variables for all records, although other "portions" are possible.) of a dataset with multiple imputations. Some

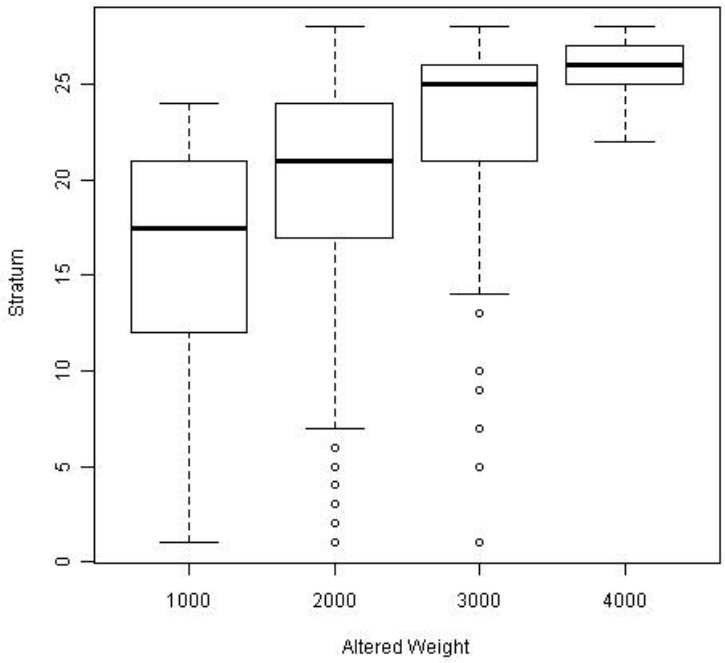

Figure 4: Plot of stratum by altered weights

risk formulations for synthetic cross-sectional data have been put forth (Reiter and Mitra, 2009; Dreschler and Reiter, 2009). However, given the lack of any widely accepted method, multiple methods, including *ad hoc* methods, have been used to illustrate disclosure protection in synthetic data applications (Kinney et al., 2010). Kinney et al. (2010) does provide a formal characterization of the disclosure protection of the synthetic LBD using the notion of differential privacy (Dwork, 2006), for a small portion of the data. Under this paradigm, the disclosure protection is provided by the (random) SDL process, and does not depend on the underlying data (Abowd and Vilhuber, 2008). A finite differential privacy bound can be computed for the maximum amount that any attacker with information about all units but one can learn about each specific unit.

Some literature on SDL for longitudinal data focuses on the additional disclosure risk presented by the longitudinal nature of the data. This may be sufficient for many databases; longitudinal data are expensive to collect, and thus often a single panel is followed for a few years before a new panel is selected, yielding only one data release per panel. Longer-term longitudinal collections, often derived from administrative data, such as the LEHD and LBD, are updated continually as new data become available. The risk to records spanning multiple data releases is not well understood.

For synthetic data, a related issue is the degree to which increasing the number of imputations increases the risk of disclosure (Reiter and Mitra, 2009). Dreschler and Reiter (2009) found in their application the disclosure risk increased with the number of imputations faster than data utility increased (based on

Karr et al. (2006a)), and struck a balance by imputing in two stages, with fewer imputations of variables considered to have higher risk.

But, releasing an updated longitudinal dataset is not the same as releasing additional imputations. The models used to generate the updated dataset may differ from the previous version, and the support of the data may change as well, particularly when entries and exits are synthesized. To illustrate, the next release of the synthetic LBD will include at least four additional years. Rather than creating synthetic data for the additional years of data and appending to the old, the entire process must be repeated on the expanded dataset. Thus every unit in the old dataset will also be in the new one, with different synthetic values.

The message is the same is in other sections: risk-utility reasoning is valuable, but falls short of truly informing or evaluating SDL decisions.

6.2 Administrative Data

There is strong and increasing pressure in the US federal statistical system and elsewhere to make greater use within surveys of administrative data. By administrative data, we mean data collected for purposes other than statistical analysis and estimation. Examples at the federal, state and local levels are income tax records, pupil information in state longitudinal data systems (SLDS) maintained by US state education authorities. (§6.1) and real property ownership records.

The reasons for using administrative data involve risk and utility indirectly if at all. These reasons are reduction of respondent burden, reduction of survey costs, and improved public relations. (Avoiding reactions of the form "Why is the government asking me what it already knows?") The envisioned mechanism for incorporating administrative into surveys is record linkage, presumably principally by means of names and addresses.

There is no convincing evidence that risk–utility paradigms are able to deal with the issues. Nor can we even be sure that the issues are at all clearly identified. Nevertheless, the ones we note next *are* issues.

The first of these is the introduction of cost, which converts two-way risk-utility tradeoffs to three-way cost–risk–utility tradeoffs. We argue here that risk-utility tradeoffs, even if understood conceptually are not broadly implemented, which is equally if not more true of cost-utility tradeoffs (Karr and Last, 2006). Nothing appears to be known about cost-risk tradeoffs, or about three-way interactions among cost, risk and utility.

Second, administrative datasets carry quality characteristics that may be unknown to, or not understood by, statistical agencies. Given that quantification of utility is not possible in highly controlled contexts, who can believe it will be possible when there is much less control?

Third, record linkage alters both risk—when it is done correctly—and utility—when it is done incorrectly. It seems difficult to believe that already fragile ways to act can accommodate these alterations.

So finally the real question may be: Are risk-utility paradigms even the "right way to think" is this setting?

7 Concluding Remarks

In some sense, our bottom line question should now be clear, and it is deeper than the one in our title: Can there be a science of data confidentiality? Implicit in this question is the belief that today there is not a science of data confidentiality. And, of course, if there is to be a science of data confidentiality, on what

in addition to risk-utility paradigms is it to be built? We challenge the data confidentiality researcher and practitioner communities to initiate a serious conversation about these issues.

Acknowledgments

This research was supported in part by NSF grants EIA–0131884 to the National Institute of Statistical Sciences (NISS) and DMS–0112069 to the Statistical and Applied Mathematical Sciences Institute. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Abowd, J. M. and Vilhuber, L. (2008). How protective are synthetic data? In Domingo-Ferrer, J. and Saygin, Y., editors, *Privacy in Statistical Databases, LNCS 5262*, pages 239–246. Springer–Verlag.

Abowd, J. M. and Woodcock, S. D. (2001). Disclosure limitation in longitudinal linked data. In Doyle, P., Lane, J. I., Theeuwes, J. J. M., and Zayatz, L. V., editors, *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, pages 215–278. Elsevier, Amsterdam.

Abowd, J. M. and Woodcock, S. D. (2004). Multiply-imputing confidential characteristics and file links in longitudinal linked data. In Domingo-Ferrer, J. and Torra, V., editors, *Privacy in Statistical Databases*. Springer–Verlag, New York.

Cox, L. H. (1980). Suppression methodology and statistical disclosure control. *J. Amer. Statist. Assoc.*, 75:337–385.

Cox, L. H. (1981). Linear sensitivity measures in statistical disclosure control. *J. Statist. Planning Inf.*, 5:152–164.

Cox, L. H. (1987a). A constructive procedure for unbiased controlled rounding. *J. Amer. Statist. Assoc.*, 82:520–524.

Cox, L. H. (1987b). A constructive procedure for unbiased controlled rounding. *J. Amer. Statist. Assoc.*, 82:520–524.

Cox, L. H. (1995). Network models for complementary cell suppression. *J. Amer. Statist. Assoc.*, 90:1453–1462.

Cox, L. H. (2003). On properties of multi-dimensional statistical tables. *J. Statist. Planning Inf.*, 17:251–273.

Cox, L. H. (2007). Contingency tables of network type: Models, Markov basis and applications. *Statistica Sinica*, 17:1371–1393.

Cox, L. H. (2008). A data quality and data confidentiality assessment of complementary cell suppression. In Domingo-Ferrer, J. and Saygin, Y., editors, *Lecture Notes in Computer Science 5262*, pages 13–23. Springer–Verlag, Heidelberg.

Cox, L. H. (2009). Vulnerability of complementary cell suppression to intruder attack. *J. Privacy and Confidentiality*, 2:235–251.

Cox, L. H. and Ernst, L. R. (1982). Controlled rounding. *INFOR: Canadian Journal of Operations Research and Information Processing*, 20:423–432.

Cox, L. H., Kelly, J. P., and Patil, R. (2004). Balancing quality and confidentiality for multivariate tabular data. In Domingo-Ferrer, J. and Torra, V., editors, *Privacy in Statistical Databases, Lecture Notes in Computer Science 3050*, pages 87–98, Berlin. Springer–Verlag.

Cox, L. H. and Kim, J. J. (2006). Effects of rounding on the quality and confidentiality of statistical data. In Domingo-Ferrer, J. and Franconi, L., editors, *Privacy in Statistical Databases 2006, Lecture Notes in Computer Science 4302*, pages 48–56, Heidelberg. Springer–Verlag.

Cox, L. H., Orelien, J. G., and Shah, B. V. (2006). A method for preserving statistical distributions subject to controlled tabular adjustment. In Domingo-Ferrer, J. and Franconi, L., editors, *Privacy in Statistical Databases 2006, Lecture Notes in Computer Science 4302*, pages 1–11, Heidelberg. Springer–Verlag.

Cox, L. H. and Zayatz, L. V. (1995). An agenda for research in statistical disclosure limitation. *J. Official Statist.*, 11(2):205–220.

De Loera, J. and Onn, S. (2004). All rational polytopes are transportation polytopes and all polytopal integer sets are contingency tables. In Bienstock, D. and Nemhauser, G., editors, *IPCO 2004, Lecture Notes in Computer Science 3064*, pages 338–351, Berlin. Springer–Verlag.

De Waal, A. G. and Willenborg, L. C. R. J. (1997). Statistical disclosure control and sampling weights. *J. Official Statist.*, 13(4):417–434.

Denogean, L. R., Karr, A. F., and Qaqish, B. F. (2007). Model-based utility of doubly random swapping. Unpublished manuscript.

Diaconis, P. and Sturmfels, B. (1998). Algebraic algorithms for sampling from conditional distributions. *Ann. Statist.*, 26:363–97.

Dobra, A. (2003). Markov bases for decomposable graphical models. *Bernoulli*, 9:1–16.

Dobra, A., Fienberg, S. E., Karr, A. F., and Sanil, A. P. (2002). Software systems for tabular data releases. *Int. J. Uncertainty, Fuzziness and Knowledge Based Systems*, 10(5):529–544.

Domingo-Ferrer, J. and í González-Nicolás (2010). Hybrid microdata using microaggregation. *Information Sciences*, 180(15):2834–2844.

Domingo-Ferrer, J., Martínez-Ballesté, A., Mateo-Sanz, J. M., and Sebé, F. (2006). Efficient multivariate data-oriented microaggregation. *The VLDB Journal*, 15:355–369.

Domingo-Ferrer, J. and Mateo-Sanz, J. M. (2002). Practical data-oriented microaggregation for statistical disclosure control. *IEEE Trans. on Knowledge and Data Engrg.*, 14(1):189–201.

Dreschler, J. and Reiter, J. P. (2009). Disclosure risk and data utility for partially synthetic data: An empirical study using the German IAB Establishment Survey. *J. Official Statist.*, 25:589–603.

Dwork, C. (2006). Differential privacy. In Bugliesi, M., Preneel, B., Sassone, V., and Wegener, I., editors, *Automata, Languages and Programming*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12. Springer–Verlag, Berlin.

Evans, T., Zayatz, L., and Slanta, D. (1998). Using noise for disclosure limitation of establishment tabular data. *J. Official Statist.*, 14:537–551.

Fellegi, I. P. (1972). On the question of statistical confidentiality. *J. Amer. Statist. Assoc.*, 67:7–18.

Fienberg, S. E. (2009). The relevance or irrelevance of weights in statistical disclosure limitation. *J. Privacy and Confidentiality*, 1(2):183–195.

Fischetti, M. and Salazar-Gonzalez, J. J. (2001). Solving the cell suppression problem on tabular data with linear constraints. *Management Sci.*, 47:1008–1026.

Fraser, B. and Wooton, J. (2005). A proposed method for confidentialising tabular output to protect against differencing. In *Monographs of Official Statistics: Work Session on Statistical Data Confidentiality*, pages 299–302, Luxembourg. Eurostat.

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statist. Sci.*, 22(1):153–164.

Gomatam, S., Karr, A. F., and Sanil, A. P. (2005). Data swapping as a decision problem. *J. Official Statist.*, 21(4):635–656.

Holan, S. H., Toth, D., Ferreira, M. A. R., and Karr, A. F. (2010). Bayesian multiscale multiple imputation with implications to data confidentiality. *J. Amer. Statist. Assoc.*, 105:564–577.

Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P., and Sanil, A. P. (2006a). A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician*, 60(3):224–232.

Karr, A. F. and Last, M. (2006). Survey costs: Workshop report and white paper. Technical Report 161, National Institute of Statistical Sciences. Available on-line at http://niss.org/sites/default/files/tr161.pdf.

Karr, A. F., Lee, J., Sanil, A. P., Hernandez, J., Karimi, S., and Litwin, K. (2001). Disseminating information but protecting confidentiality. *IEEE Computer*, 34(2):36–37.

Karr, A. F., Sanil, A. P., and Banks, D. L. (2006b). Data quality: A statistical perspective. *Statistical Methodology*, 3(2):137–173.

Kelly, J., Golden, B., and Assad, A. (1992). Cell suppression: Disclosure protection for sensitive tabular data. *Networks*, 22:397–417.

Kinney, S. K., Reiter, J. P., Reznek, A. P., Miranda, J., Jarmin, R. S., and Abowd, J. M. (2010). Toward unrestricted public-use business microdata: The Longitudinal Business Database. Technical report, National Institute of Statistical Sciences.

Kuhn, T. (1962). *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago, IL.

Little, R. J. A. (2003). To model or not to model? competing modes of inference for finite population sampling. Technical report, University of Michigan School of Public Health. Available on-line at http://www.bepress.com/umichbiostat/paper4/.

Lu, W. W. and Sitter, R. R. (2008). Disclosure risk and replication-based variance estimation. *Statistica Sinica*, 18:1669–1687.

Mitra, R. and Reiter, J. P. (2006). Adjusting survey weights when altering identifying design variables via synthetic data. In Domingo-Ferrer, J. and Franconi, L., editors, *Privacy in Statistical Databases 2006*, volume 4302 of *Lecture Notes in Comp. Sci.*, pages 177–188. Springer–Verlag, Berlin/Heidelberg.

National Institute of Statistical Sciences (2003). NISS Data Swapping Toolkit User Documentation. Available on-line at www.niss.org/software/dstk.html.

National Research Council Committee on National Statistics (2000). Improving access to and confidentiality of research data: Report of a workshop.

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

Oganian, A. and Karr, A. F. (2006). Combinations of SDC methods for microdata protection. In Domingo-Ferrer, J. and Franconi, L., editors, *Privacy in Statistical Databases 2006*, volume 4302 of *Lecture Notes in Comp. Sci.*, pages 102–113. Springer–Verlag, Berlin/Heidelberg.

Park, I., Dohrmann, S., Montaquila, J., Mohadjer, L., and Curtin, L. R. (2006). Reducing the risk of data disclosure through area masking: Limiting biases in variance estimation. In *Proc. ASA Section on Physical and Engineering Sciences*, pages 1761–1767, Alexandria, VA. American Statistical Association.

Pistone, G., Riccomagno, E., and Wynn, H. (2001). Algebraic statistics: Computational commutative algebra in statistics. In Bunea, F., Isham, V., Keiding, N., Louis, T. A., Smith, R. L., and Tong, H., editors, *Monographs on Applied Statistics and Probability 89*. Chapman and Hall, London.

Reiter, J. P. and Mitra, R. (2009). Estimating risks of identification disclosure in partially synthetic data. *J. Privacy and Confidentiality*, 1:99–110.

Schlörer, J. (1980). Disclosure from statistical databases: Quantitative aspects of trackers. *ACM Trans. Database Systems*, 5:467–492.

Templ, M. (2008). Statistical disclosure control for microdata using the R-package sdcMicro. *Trans. Data Privacy*, 1(2):67–85.

Trottini, M. (2001). A decision-theoretic approach to data disclosure problems. *Res. Official Statist.*, 4:7–22.

Trottini, M. (2003). *Decision Models for Data Disclosure Limitation*. PhD thesis, Carnegie Mellon University. Available on-line at www.niss.org/dgii/TR/Thesis-Trottini-final.pdf.

Willenborg, L. C. R. J. and de Waal, T. (1996). *Statistical Disclosure Control in Practice*. Springer–Verlag, New York.

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

Woo, M.-J., Reiter, J. P., and Karr, A. F. (2008). Estimation of propensity scores using generalized additive models. *Statistics in Medicine*, 23:3806–3816.

Woo, M.-J., Reiter, J. P., Oganian, A., and Karr, A. F. (2009). Global measures of data utility for microdata masked for disclosure limitation. *J. Privacy and Confidentiality*, 1(1):111–124.

Zayatz, L. V., Massell, P., and Steel, P. (1999). Disclosure limitation practices and research at the Census Bureau. *Netherlands Official Statistics*, 14:26–29.