NISS

Data Swapping as a Decision Problem

Shanti Gomatam, Alan F. Karr and Ashish P. Sanil

Technical Report Number 140
revised October, 2004

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

Data Swapping as a Decision Problem

Shanti Gomatam*, Alan F. Karr and Ashish P. Sanil
National Institute of Statistical Sciences
Research Triangle Park, NC 27709–4006, USA
{sgomatam,karr,ashish}@niss.org

October 25, 2004

Abstract

We construct a decision-theoretic formulation of data swapping in which quantitative measures of disclosure risk and data utility are employed to select one release from a possibly large set of candidates. The decision variables are the swap rate, swap attribute(s) and possibly, constraints on the unswapped attributes. Risk–utility frontiers, consisting of those candidates not dominated in (risk, utility) space by any other candidate, are a principal tool for reducing the scale of the decision problem. Multiple measures of disclosure risk and data utility, including utility measures based directly on use of the swapped data for statistical inference, are introduced. Their behavior and resulting insights into the decision problem are illustrated using data from the Current Population Survey, the well-studied "Czech auto worker data" and data on schools and administrators generated by the National Center for Education Statistics.

1 Introduction

Data swapping (Gomatam and Karr, 2003; Willenborg and de Waal, 1996, 2001) is a technique for statistical disclosure limitation that works at the microdata (individual data record) level. Confidentiality protection is achieved by selectively modifying a fraction of the records in the database by exchanging a subset of attributes between selected pairs of records. Data swapping makes it impossible for an intruder to be certain of having identified an individual or entity in the database, because no record is certain to be unaltered.

Data swapping is of course, not new. The seminal papers on the subject is Dalenius and Reiss (1982) and Reiss (1984), and recent references include Fienberg and McIntyre (2004). A formal definition (Willenborg and de Waal, 2001) uses elementary swaps. An *elementary swap* is a selection of two records from the microdata and an interchange of the values of attributes being swapped for these two records. When the candidates for each swap pair are picked at random we will refer to

*Currently at US Food and Drug Administration, Rockville, MD.

the resulting swaps as *random swaps*. We assume that elements of a swap pair are picked without replacement, so that no record appears in more than one swap pair. We also allow only *true swaps*, in the sense that both the swap attribute and at least one unswapped attribute must differ.[1] For multiple swap attributes, *all* attributes are swapped simultaneously, and all swap attributes must differ. The algorithm to perform the swapping is described in Appendix A and Sanil et al. (2003).

In the past, implementation of data swapping by statistical agencies has been a matter of judgement. The US Census Bureau is a leading user of data swapping, especially when the "swap attribute" (see below) is geography; this special case is sometimes termed "switching" (Cox and Zayatz, 1993). Agency behavior is typically conservative, erring on the side of too much protection of confidentiality rather than risking too little. Moreover, compared to immense attention to the effects of data swapping on confidentiality, much less attention has been paid to the effects of data swapping on the usefulness of the released data. Clearly data swapping distorts the data: joint distributions involving both swapped and unswapped attributes change. This decreases the value of the data for purposes such as statistical inference. Confidentiality protection and data utility must be traded off: they are, in economic terminology, *substitutes*—more of one entails less of the other.

In this paper, we formulate implementation of data swapping as a decision problem with explicit tradeoff of quantified measures of disclosure risk and data utility. In its simplest form, this problem entails selection of one or more swap attributes and the *swap rate*, the fraction of records for which swapping occurs. More complex versions of the problem allow constraints on unswapped attributes. For example, an unswapped attribute may be forced to remain unchanged—preventing swapping across geographical boundaries, for example—or forced to change.

Our formulation of data swapping as a decision problem appears in §2, together with two complementary approaches to solving the problem. In §3 and 4 we introduce particular measures of disclosure risk and data utility, the latter conceptualized in part as lack of data distortion. These are illustrated using example data from the Current Population Survey (CPS) (Census Bureau, 2002). In §5 we describe risk–utility tradeoffs for three databases—CPS data, data on school administrators from the National Center for Education Statistics (NCES), and the Czech automobile worker database (Edwards and Havraneek, 1985); §6 contains a concluding discussion.

2 Problem Formulation

In this section, we formulate data swapping as a decision problem: what must be decided (§2.1) and how quantified measures of disclosure risk and data utility (§2.2) facilitate solution of the problem (§2.3).

A number of model-based frameworks for trading off risk and utility have been proposed (Duncan et al., 2004, 2001; Trottini, 2001, 2003; Zaslavsky and Horton, 1998). The terminology "R–U

[1] In some early versions of our software (Sanil et al., 2003), a looser definition of "true swap" was employed, which required that each record change, but not that the database change. For example, with Age with swap attribute, (Age = $\geq$ 50, Sex = Male) $\leftrightarrow$ (Age $<$ 50, Sex = Male) would have been a true swap under the earlier formulation, but no longer is one.

confidentiality map" for this tradeoff is employed in Duncan et al. (2004) in the context of top-coding and in Kim and Winkler (1995) to denote a simulation experiment for perturbed, by addition of noise, multivariate data. A Bayesian approach to contrasting risk and utility for cell suppression is studied in Zaslavsky and Horton (1998). A risk–utility approach statistical disclosure limitation for tabular data, in which releases are marginal subtables of a large contingency table, appears in Dobra et al. (2002), Dobra et al. (2003) and Karr et al. (2003). We do not build specifically on any of these, but our approach is clearly in the same spirit.

2.1 Structure of the Decision Problem

Consider a database $\mathcal{D}$ consisting of a single table of data records having only categorical attributes. Much of the formulation in this paper but fewer of the specifics such as measures of disclosure risk and data distortion, generalizes to "continuous" attributes.

The decision problem for data swapping involves three principal stages.[2]

The first stage is to decide whether to use data swapping at all, and whether to use data swapping alone or in conjunction with other strategies for statistical disclosure limitation. This choice lies largely outside the realm of this paper, and may be dictated by agency practice, political issues or scientific considerations. In general, data swapping is used in situations where the release of altered microdata is preferred to that of (possibly exact) summaries or analyses of the data.

Extensions of our risk–utility paradigm may allow quantification of tradeoffs among multiple statistical disclosure limitation strategies, although clearly additional research is required before this becomes a reality.

Second, if data swapping *is employed*, disclosure risk and data utility measures must be selected, which are used as shown in §2.3 to perform the third stage of the decision process. Examples of such measures appear in §3 and 4.

Third, the release must be selected from some set $\mathcal{R}_{\text{cand}}$ of candidate releases, which ordinarily entails choosing the

Swap rate, the fraction of records in the database $\mathcal{D}$ for which swapping will occur.

Swap attributes, those attributes whose values are exchanged between randomly selected pairs of records in $\mathcal{D}$.

Constraints on the unswapped attributes, which are optional. Such constraints may require or forbid equality of unswapped attributes.

More specifically, as in §2.2, candidate releases are parameterized by a swap rate, the swap attributes and constraints, and constructed by actually performing the swap. Then, values of disclosure risk and data utility are computed for each candidate release, and used to select which candidate to release.

[2] In effect, one decision precedes all of these: to release microdata at all, as opposed to summaries or statistical analyses of the data. As more external databases become available and record linkage technologies improve, any useful release of microdata may be too threatening to confidentiality. An initial look at a "world without microdata" appears in Gomatam et al. (2004).

Although in principle the risk–utility paradigm in §2.2–2.3 can be used to select all three of these, we envision that it will be used frequently to select swap attributes, less frequently to select the swap rate, and only rarely to select the constraints. Ordinarily, constraints would be imposed exogenously on the basis of domain knowledge. For example, it may be declared that swapping may not occur across state lines, because doing so would lead to released microdata that are inconsistent with state-level totals available elsewhere. Similarly, constraints may be necessary to prevent physically infeasible (and hence detectable) swapped records, such as males who have undergone hysterectomies. Even in such cases, however, our methods can still be used to evaluate the impact of the constraints on disclosure risk and data utility.

2.2 Mathematical Representation

Let d be the number of (categorical) attributes in the *pre-swap* database $\mathcal{D}_{\text{pre}}$. The mathematical abstraction of the decision problem laid out in §2.1 entails specification of candidate releases, a disclosure risk measure and a data utility measure.

Releases. We parameterize candidate releases as

$$R = (r, \text{AS}_1, \ldots, \text{AS}_d), \tag{1}$$

where r is the swap rate, and for each attribute i, the attribute specification

$$\text{AS}_i \in \{\text{S, F, C, U}\} \tag{2}$$

determines whether attribute i is swapped (S), must remain fixed (F), must change (C), or is neither swapped nor constrained (U).

Release Space. Because in practice only finitely many swap rates are consider, the *release space* $\mathcal{R}$ is finite, but may be very large. Even for a fixed swap rate, there are on the order of $4^{d-1/2}$ possible releases, corresponding to all possible combinations of S, F, C and U in (2) other than $(C, \ldots, C)$, $(F, \ldots, F)$, $(S, \ldots, S)$ and $(U, \ldots, U)$ and accounting for complementarity—swapping one set of attributes is equivalent to swapping its complement.

Candidate Release Space. In many settings, therefore, it is convenient or necessary to consider a smaller set $\mathcal{R}_{\text{cand}}$ of *candidate releases*. For example, in §5.1, where $d = 8$, there are 108 candidate releases corresponding to three swap rates, all possible one- and two-attribute swaps, and no constraints.

Note that candidate releases correspond to parameterized rather than actual releases. For each release $R \in \mathcal{R}_{\text{cand}}$ we construct an actual release—a post-swap database $\mathcal{D}_{\text{post}}(R)$, using the algorithm in Appendix A. Define the actual candidate release space

$$\mathcal{R}^{\text{act}}_{\text{cand}} = \left\{ \mathcal{D}_{\text{post}}(R) : R \in \mathcal{R}_{\text{cand}} \right\}, \tag{3}$$

one of whose elements will be released. The selection problem is to choose *which one*. Its solution, which we describe in §2.3, requires quantified measures of disclosure risk and data utility; specific examples for data swapping are presented in §3–4.

Because the data swapping algorithm in Appendix A entails randomization, there is ambiguity in (3): different choices of the randomization seed yield different post-swap databases, even for the same parameterized release R in (1). It is even possible to include the randomization seed in the choice problem, but for simplicity we do not. In fact, when there is little possibility of confusion, we treat $R \in \mathcal{R}_{\text{cand}}$ and $\mathcal{D}_{\text{post}}(R) \in \mathcal{R}^{\text{act}}_{\text{cand}}$ as synonymous.

Disclosure Risk. The *disclosure risk measure* is a function $\mathbf{DR} : \mathcal{R} \to \mathbb{R}$ with the interpretation that $\mathbf{DR}(R)$ is the disclosure risk associated with the release R.[3] If $\mathcal{R}_{\text{cand}}$ is immutable, then of course **DR**, as well as the data utility measure **DU**, need only be defined on it, and not necessarily on all of $\mathcal{R}$. The disclosure risk function need not have any particular properties other than sensibly abstracting disclosure risk. However, in settings such as tabular data, in which the release space is partially ordered, the disclosure risk measure must be monotone with respect to the partial order.

Data Utility. The *data utility measure* is a function $\mathbf{DU} : \mathcal{R} \to \mathbb{R}$ with the interpretation that $\mathbf{DU}(R)$ is the utility of the release R.

2.3 Solution of the Decision Problem

Given disclosure risk and data utility measures, the data swapping decision problem can be solved in two distinct ways.

Utility Maximization. In this case, the optimal release R^* is chosen that maximizes data utility subject to an upper bound constraint on disclosure risk:

$$\begin{aligned} R^* &= \arg\max_{R \in \mathcal{R}_{\text{cand}}} \mathbf{DU}(R) \\ &\text{s.t. } \mathbf{DR}(R) \le \alpha, \end{aligned} \tag{4}$$

where α is the bound on disclosure risk, which must be specified by the decision maker.

Risk–Utility Frontiers. Especially but not only if $\mathcal{R}_{\text{cand}}$ is small, then it may be more insightful simply to compare releases R in terms of risk and utility simultaneously, using the partial order $\preceq_{\text{RU}}$ defined by

$$R_1 \preceq_{\text{RU}} R_2 \Leftrightarrow \mathbf{DR}(R_2) \le \mathbf{DR}(R_1) \qquad \text{and} \qquad \mathbf{DU}(R_2) \ge \mathbf{DU}(R_1). \tag{5}$$

If $R_1 \preceq_{\text{RU}} R_2$, then clearly R_2 is preferred to R_1 because it has both lower disclosure risk and higher greater utility. Only elements of $\mathcal{R}_{\text{cand}}$ on the *risk–utility frontier* $\partial\mathcal{R}_{\text{cand}}$ consisting of the maximal elements of $\mathcal{R}_{\text{cand}}$ with respect to the partial order (5) need be considered further. Ordinarily, as illustrated schematically in Figure 1 and for real data in Figures 3 and 4, the frontier is much smaller than $\mathcal{R}_{\text{cand}}$. Calculation of the frontier can be done using existing algorithms for finding the maxima in a set of vectors (Kung et al., 75). These algorithms have a worst case complexity of $\mathrm{O}(N \log N)$, where $N = \#\{R_{\text{cand}}\}$. However, the average case complexity is $\mathrm{O}(N)$

[3] This is an example of the simplification from the preceding paragraph. Strictly speaking, disclosure risk is a function of $\mathcal{D}_{\text{post}}(R)$ rather than R, and indeed, the examples in §3 show this.

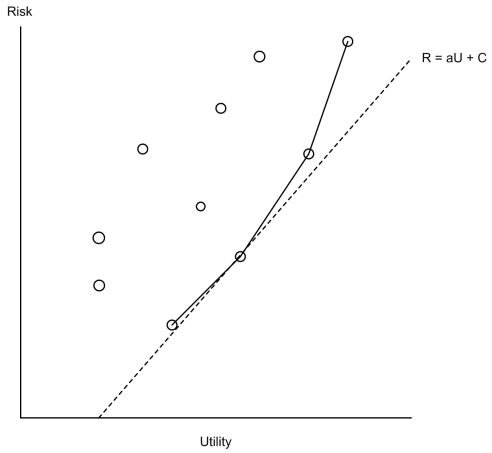

Figure 1: Conceptual risk-utility frontier and optimal release for a linear tradeoff between risk and utility.

for a large class of distributions of the data (Bentley et al., 1978). In any event, computation of the frontier only comprises a small part of the overall computational effort.

Selection of a release on the risk–utility frontier can be done by assessing the risk–utility balance subjectively or quantitatively, by means of an objective function that relates risk and utility. To illustrate, the dashed line in Figure 1 corresponds to a linear risk-utility relationship of the form

$$\mathbf{DR} = a \times \mathbf{DU} + c,$$

and the figure identifies the release on $\partial\mathcal{R}_{\text{cand}}$ that is optimal for a particular value of a. Similar approaches have been used in economics to maximize consumer utility for the purchase of a combination of two commodities.

Risk–utility frontiers also facilitate solution of the utility maximization problem (4), because the optimal release R^* must lie on the frontier.

3 Disclosure Risk Measures

Here we describe two disclosure risk measures that are both derived from the concept that re-identification of data subjects is the primary threat to confidentiality.

3.1 Small Cell Counts

Especially for census data, population uniques or near uniques are potentially riskier than other elements. For categorical data, these elements are contained in small count cells in the contingency table created by using all attributes in the data.

The n-rule, which is widely used in statistical disclosure limitation (Willenborg and de Waal, 2000), considers records that fall in cells with count (strictly) less than n (typically $n = 3$) to be at

risk. Reflecting this, we define risk as the proportion of unswapped records in small count cells in the table created from the post-swap data:

$$\mathbf{DR}(R) = \frac{\sum_{C_1,C_2} \text{Number of unswapped records in } \mathcal{D}_{\text{post}}(R)}{\text{Total number of unswapped records in } \mathcal{D}_{\text{post}}(R)}, \tag{6}$$

where C_1 and C_2 are the cells in the full data table associated with $\mathcal{D}_{\text{post}}(R)$ with counts of 1 and 2 respectively. For survey data such measures are well-known to be extremely conservative.

Unlike the data distortion measures in §4, which are stated for categorical data but generalize readily to continuous data, the disclosure measure of (6) makes sense only for categorical data.

3.2 Record Linkage

A number of authors (Cox, 1979; Domingo-Ferrer et al., May, 2001; Lambert, 1993; Spruill, 1982; Winkler, 1998; Willenborg and de Waal, 2001; Yancey et al., 2002) have considered disclosure risk measures based on re-identification through record linkage.

For example, let $\mathcal{D}_{\text{ext}}$ be an external database containing attributes in common with $\mathcal{D}$ (and the same attributes in common with any $\mathcal{D}_{\text{post}}(R)$), and for each record $r \in \mathcal{D}_{\text{post}}(R)$ let $n(r) = n(R; r)$ be the number of records in $\mathcal{D}_{\text{ext}}$ that agree with r on the common attributes. These are candidates for linkage to r. For purposes of statistical disclosure limitation, larger values of $n(r)$ are better, because they make record linkage more uncertain. A disclosure risk measure that captures this is

$$\mathbf{DR}'(R) = \frac{\text{Number of records in } \mathcal{D}_{\text{post}}(R) \text{ with } n(r) \le \beta}{\text{Total number of records in } \mathcal{D}_{\text{post}}(R)}, \tag{7}$$

where β is a threshold.

4 Data Utility Measures

Let $\mathcal{D}_{\text{pre}}$ denote the database prior to swapping, and let $\mathcal{D}_{\text{post}}(R)$ denote the post-swap database for candidate release R. In this section, we describe two classes of data utility measures that capture the extent to which $\mathcal{D}_{\text{post}}(R)$ differs from $\mathcal{D}_{\text{pre}}$. The first of these (§4.1) measures explicitly the distortion introduced by data swapping. Distortion is data *disutility*, so that if **DD** is a measure of data distortion, then $\mathbf{DU} = -\mathbf{DD}$ is the associated measure of data utility.

Direct measures of distortion are general but blunt. They are disconnected from specific uses of the data, such as statistical inference. In §4.2 we present data utility measures that quantify the extent to which inferences (in our case, using log-linear models) based on $\mathcal{D}_{\text{post}}(R)$ differ from those based on $\mathcal{D}_{\text{pre}}$.

4.1 Data Distortion

Recall that the data are categorical. Our data distortion measures are based on viewing $\mathcal{D}_{\text{pre}}$ and $\mathcal{D}_{\text{post}}(R)$ as contingency tables, and thus (when normalized) as distributions on the space $\mathcal{I}$ that

indexes cells in these tables. Mathematically, $\mathcal{I}$ is the Cartesian product of the sets of category values for each attribute. We let $\mathcal{D}_{\text{pre}}(c)$ be the cell count associated with cell $c \in \mathcal{I}$.

The distortion measures all have the form

$$\mathbf{DD}(R) = d(\mathcal{D}_{\text{pre}}, \mathcal{D}_{\text{post}}(R)), \tag{8}$$

where d is a metric on an appropriate space of distributions. Recall also that data swapping changes only joint distributions of the attributes that involve both swap attributes and unswapped attributes. Distortion measures of the form (8) involve *all attributes*.

Hellinger distance (Le Cam and Yang, 1990) is given by

$$\text{HD}(\mathcal{D}_{\text{pre}}, \mathcal{D}_{\text{post}}(R)) = \frac{1}{\sqrt{2}} \sqrt{\sum_{c \in \mathcal{I}} \left(\sqrt{\mathcal{D}_{\text{pre}}(c)} - \sqrt{\mathcal{D}_{\text{post}}(R, c)} \right)^2}. \tag{9}$$

Note that the same absolute difference between $\mathcal{D}_{\text{pre}}(c)$ and $\mathcal{D}_{\text{post}}(R, c)$ affects the Hellinger distance to a greater extent when the value of $\mathcal{D}_{\text{pre}}(c)$ is small. Hellinger distance also corresponds to Cressie–Read divergence (Cressie and Read, 1988) with $\lambda = -0.5$.

Total variation distance is given by

$$\text{TV}(\mathcal{D}_{\text{pre}}, \mathcal{D}_{\text{post}}(R)) = \frac{1}{2} \sum_{c \in \mathcal{I}} \left| \mathcal{D}_{\text{pre}}(c) - \mathcal{D}_{\text{post}}(R, c) \right|. \tag{10}$$

Entropy change is based on Shannon entropy, which for $\mathcal{D}_{\text{pre}}$ is given by

$$h(\mathcal{D}_{\text{pre}}) = -\sum_{c \in \mathcal{I}} \mathcal{D}_{\text{pre}}(c) \log \left[\mathcal{D}_{\text{pre}}(c) \right],$$

and is conventionally interpreted at the amount of uncertainty in $\mathcal{D}_{\text{pre}}$. Entropy change, then, constitutes another measure of data distortion:

$$\Delta h(\mathcal{D}_{\text{pre}}, \mathcal{D}_{\text{post}}(R)) = h(\mathcal{D}_{\text{post}}(R)) - h(\mathcal{D}_{\text{pre}}). \tag{11}$$

Positive values of $\Delta h(\mathcal{D}_{\text{pre}}, \mathcal{D}_{\text{post}}(R))$ indicate that swapping has increased the uncertainty in the data. Related distortion measures involving conditional entropy have also been employed (Willenborg and de Waal, 2001).

We illustrate these measures using an 8-attribute database CPS-8D extracted from the 1993 CPS. The attributes, abbreviations we use for them and category values appear in Table 1. There are 48,842 data records; the associated full table contains 2880 cells, of which 1695 are non-zero. In reality, the fact that we have survey rather than census data would represent additional protection against disclosure.

Figure 2 shows the values of $\text{HD}(\mathcal{D}_{\text{pre}}, \mathcal{D}_{\text{post}}(R))$, $\text{TV}(\mathcal{D}_{\text{pre}}, \mathcal{D}_{\text{post}}(R))$ and $\Delta h(\mathcal{D}_{\text{pre}}, \mathcal{D}_{\text{post}}(R))$ for the CPS-8D data for 24 candidate releases corresponding to swap rates of 1%, 5% and 10% and all single-attribute swaps. (These and other results in this paper were produced using—in

Attribute Name (*ShortName*)	Abbreviation	Categories
Age (in years) (*Age*)	A	$<$25, 25–55, $>$55
Employer Type (*EmpTyp*)	W	Govt., Priv., Self-Emp., Other
Education (*Edu*)	E	$<$HS, HS, Bach, Bach+, Coll
Marital Status (*MS*)	M	Married, Other
Race (*Race*)	R	White, Non-White
Sex (*Sex*)	S	Male, Female
Average Weekly Hours Worked (*AvgHrs*)	H	< 40, 40, > 40
Annual Salary (*AnnSal*)	I	$<$\$50K, \$50$K+$

Table 1: Attributes and attribute categories for the CPS-8D data. The short names are used only in Figure 2
and the text, including Appendix B. The abbreviations appear in Figure 4.

this case a prototype of—the NISS Data Swapping Toolkit (National Institute of Statistical Sciences, 2003a).) As expected, distortion increases as the swap rate increases, approximately linearly. Figure 2 shows rather dramatically that swapping some attributes induces more distortion than swapping others, an issue that we discuss at greater length in §5. In general, though, the three distortion measures track each other very closely, and in particular, total variation distance and entropy change result in almost the same ordering of swap variables. Hellinger distance shows a somewhat different ordering, to which *Age* and *AvgHrs* appear to contribute the most.

Additional data distortion measures that are restricted to two-attribute databases (or more generally, if only distortion of bivariate distributions is of interest), appear in Appendix B.

4.2 Inference-Based Measures of Utility

As noted in the lead-in to this section, data distortion is a blunt measure of data utility because it does not address directly inferences that are drawn from the post-swap data. There is, of course, indirect information, because nearly all inference procedures are in some sense "continuous" with respect to the data, so that low distortion implies nearly correct inference. Here, by contrast, we describe data utility measures that account explicitly for inference in the form of log-linear models (Bishop et al., 1975) of the data.

Let $\mathbf{M}^* = \mathbf{M}^*(\mathcal{D}_{\text{pre}})$ be the "optimal" log-linear model of the pre-swap database $\mathcal{D}_{\text{pre}}$, according to some criterion, for example, the Akaike information criterion (AIC) (Akaike, 1973) or Bayes information criterion (BIC) (Schwarz, 1978). Concretely, $\mathbf{M}^*$ can be thought of in terms of its minimal sufficient statistics, that is, the set of marginal subtables of the contingency table associated with $\mathcal{D}_{\text{pre}}$ representing the highest-order interactions present. Let $\mathcal{L}_{\mathbf{M}^*}(\cdot)$ be the log-likelihood function associated with $\mathbf{M}^*$. Then as measure of data utility we employ the log-likelihood ratio

$$\mathbf{DU}_{\text{llm}}(R) = \mathcal{L}_{\mathbf{M}^*}(\mathcal{D}_{\text{post}}(R)) - \mathcal{L}_{\mathbf{M}^*}(\mathcal{D}_{\text{pre}}); \tag{12}$$

the llm subscript abbreviates "log-linear model." Although in general $\mathbf{DU}_{\text{llm}}(R) < 0$ in (12),

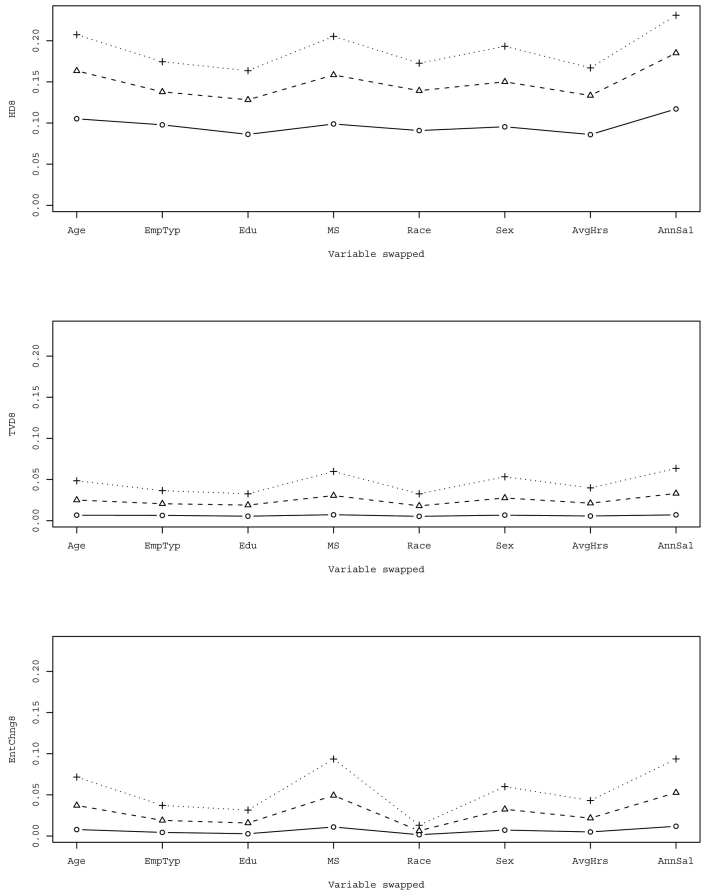

Figure 2: Graph of Hellinger (top) and total variation (middle) distances and entropy change (bottom) for 1% swap ($\bigcirc$), 5% ($\triangle$) and 10% swap rates ($+$).

because of the randomization in data swapping, this is not a logical necessity.

The rationale is that higher values of $\mathbf{DU}_{\text{llm}}(R)$ indicate that $\mathbf{M}^*$ remains a good model for $\mathcal{D}_{\text{post}}(R)$. This is not, however, completely equivalent to saying that the same inferences would be drawn from $\mathcal{D}_{\text{post}}(R)$ as from $\mathcal{D}_{\text{pre}}$, since data users do not have access to $\mathbf{M}^*$. A more complex inference-based measure of utility might, for example, compare $\mathbf{M}^*$ to a similarly optimal model $\mathbf{M}^*(\mathcal{D}_{\text{post}}(R))$ of the post-swap data. Precisely how to do so, however, requires further research. One example would be whether $\mathbf{M}^*(\mathcal{D}_{\text{pre}})$ and $\mathbf{M}^*(\mathcal{D}_{\text{post}}(R))$ have the same minimal sufficient statistics, but this measure is highly discontinuous.

In fact, Fienberg et al. (1998) illustrate a systematic procedure of perturbation (swapping) of contingency table entries so that the margins corresponding to the minimal sufficient statistics of $M^\star$ are preserved. Their procedure entails the computationally intensive task of computing "Gröbner bases" corresponding to the marginals and it does not scale well to higher dimensional tables.

				B	no		yes	
F	E	D	C	A	no	yes	no	yes
neg	< 3	< 140	no		44	40	112	67
			yes		129	145	12	23
		≥ 140	no		35	12	80	33
			yes		109	67	7	9
	≥ 3	< 140	no		23	32	70	66
			yes		50	80	7	13
		≥ 140	no		24	25	73	57
			yes		51	63	7	16
pos	< 3	< 140	no		5	7	21	9
			yes		9	17	1	4
		≥ 140	no		4	3	11	8
			yes		14	17	5	2
	≥ 3	< 140	no		7	3	14	14
			yes		9	16	2	3
		≥ 140	no		4	0	13	11
			yes		5	14	4	4

Table 2: The Czech automobile worker database (Edwards and Havraneek, 1985). High-risk cells are shown by boxes.

Also, the case when the optimal set of marginals is itself too risky to release is still unresolved. However, the Fienberg *et al.* strategy is a complementary approach with similar goals which, when computationally feasible, does provide superior information for inference to sophisticated users—those capable of running extensive Markov chain Monte Carlo simulations using the Gröbner bases employed to carry out the swapping.

In §5.2 we illustrate $\mathbf{DU}_{\text{llm}}$ for the "Czech auto worker data," an intensively studied (Edwards and Havraneek, 1985; Dobra et al., 2002), 6-attribute database containing risk factors for coronary thrombosis for 1841 Czechoslovakian automobile factory workers who took part in a prospective epidemiological study. The associated contingency table, which contains $2^6 = 64$ cells and is not sparse, appears in Table 2. The six dichotomous attributes are defined as follows: A indicates whether the worker "smokes," B corresponds to "strenuous mental work," C corresponds to "strenuous physical work," D corresponds to "systolic blood pressure," E corresponds to "ratio of β and α lipoproteins," and F represents "family anamnesis of coronary heart disease." There are three high risk cells, one with count 1 and two with count two.

5 Risk–Utility Tradeoffs

In this section, we illustrate risk–utility tradeoffs for a variety of databases and utility measures: the CPS-8D database (§5.1), school administrator data from the NCES (§5.3) and the Czech auto-

mobile worker database of Table 2 (§5.2). Rather than a "full factorial" design of all risk measures and all utility measures on each database, we report selected results that illuminate our risk–utility methodology.

5.1 CPS-8D Data

Here we illustrate risk–utility tradeoffs for the CPS-8D data for a candidate release space $\mathcal{R}_{\text{cand}}$ containing 108 cases corresponding to candidate releases comprising all (8) single-attribute swaps and all (28) two-attribute swaps together with swap rates of 1%, 2% and 10% of the data. The disclosure risk measure is given by (6) and data utility is derived from Hellinger distance-measured distortion:

$$\mathbf{DU}(R) = -\mathbf{DD}(R) = -\text{HD}\left(\mathcal{D}_{\text{pre}}, \mathcal{D}_{\text{post}}(R)\right).$$

The results, which were obtained using the NISS Data Swapping Toolkit, are shown separately for each of the three swap rates in Figure 3, with the swap attributes identified, and with all three rates on one plot in Figure 4. Since $\mathbf{DU} = -\mathbf{DD}$, these plots are reversed left–to–right as compared to Figure 1, and $\partial\mathcal{R}_{\text{cand}}$ is now the "southwest boundary."

In Figure 3, lines connect the cases on the frontier $\partial\mathcal{R}_{\text{cand}}$. A user who has already decided on a rate need only look at the plot corresponding to that rate and make a decision as to which candidate release on the frontier best captures the relevant risk and utility tolerances. For example, and restricting attention to the 2% rate, if the optimization criterion of (4) were employed, which in this case translates to

$$\begin{aligned} R^* &= \arg\min_{R\in\mathcal{R}_{\text{cand}}} \mathbf{DD}(R) \\ &\text{s.t. } \mathbf{DR}(R) \le \alpha, \end{aligned} \tag{13}$$

and if $\alpha = .014$, then the optimal release corresponds to swap attributes *Sex* and *EmpTyp*, which is labeled by "WS" in the middle panel of Figure 3.

Alternatively, a user who is undecided about the swap rate would select from the combined frontier $\partial\mathcal{R}_{\text{cand}}$ generated by putting together all swaps for the rates of interest, as in Figure 4. The frontier for the combined plot is a strict subset of the union of the three individual frontiers. For example, the 10% swap of *Educ*, which was on the frontier for the 10% swap rate, is dominated by many 1% and 2% swaps.

Figure 4 also clearly illustrates how distortion increases and risk decreases with increasing swap rate. Single-attribute swaps tend to be riskier than two-attribute swaps but show less mean distortion than two-attribute swaps. As swap rate increases, variability in both risk and Hellinger distance increases.

5.2 Czech Automobile Worker Data

The log-linear model based data utility measure $\mathbf{DU}_{\text{llm}}(R)$ in (12) was calculated for the Czech automobile worker data in Table 2 for 21 releases corresponding to all one- and two-attribute swaps, with a single swap rate of 10%, with the "batch swap" capability (National Institute of Statistical Sciences, 2003b) of the NISS Data Swapping Toolkit used to perform the swapping. The

optimal model $\mathbf{M}^* = \mathbf{M}^*(\mathcal{D}_{\text{pre}})$ under either AIC or BIC has as sufficient statistics the marginal subtables

$$\Big\{[ABCD], [ADE], [FB]\Big\}. \tag{14}$$

This model is also well-recognized as the "best" model on the basis of domain knowledge (Edwards and Havraneek, 1985; Whittaker, 1990).

Figure 5 shows the associated risk–utility plot, with risk given by (6). Points there are labeled by swap attributes, with A, . . ., F the single-attribute swaps and fe, . . ., ba the two-attribute swaps. Since this is a risk–utility (not risk–distortion) plot, it is comparable to Figure 1. The frontier is the southeast boundary of the set of candidate releases:

$$\partial\mathcal{R}_{\text{cand}} = \Big\{\mathsf{b}, \mathsf{ed}, \mathsf{fe}, \mathsf{ec}, \mathsf{fa}, \mathsf{fd}\Big\}.$$

In Figure 5, the points fc and fb are clearly anomalous: they have extremely low utility. One, but only one, of these corresponds to a marginal in (14).

One obvious question is whether the inference-based utility measure $\mathbf{DU}_{\text{llm}}$ actually "picks up" some sort of signal that is obscured by the (general but as we termed it "blunt") Hellinger distance data distortion measure of (9). Figure 6 plots ($\mathbf{DU}_{\text{llm}}(R)$, HD) pairs for the same 21 cases appearing in Figure 5. The relationship is ambiguous at best, which we interpret as meaning that $\mathbf{DU}_{\text{llm}}(R)$ and HD are indeed different. Indeed, ignoring the anomalous points fc and fb, there seems to be little apparent relationship between $\mathbf{DU}_{\text{llm}}(R)$ and HD.

5.3 NCES Data

Here we illustrate insights produced by our decision-theoretic formulation of data swapping, using data from the NCES. Specifically, we use eight categorical attributes extracted from the 1993 Common Core of Data (CCD) Public Elementary/Secondary School Universe Survey data file and the 1993–94 Schools and Staffing Survey (SASS) Public and Private Administrator data file. The attribute names and category values appear in Table 3.

Figure 7 shows the results of 800 swaps of the NCES data, corresponding to 100 realizations each of the eight one-attribute swaps. The realizations differ only by the initial seed of the random number generator used to perform the choices of swap pairs (see Appendix A). The swap rate in all cases is 10%. In each panel of Figure 7, the 100 cases involving a particular attribute are highlighted.

"Administrator Experience" is highlighted in the upper left panel, giving first a visual expression of the "random variability" inherent in the swapping algorithm, which we interpret as non-trivial but not dramatic. Perhaps more important, this panel demonstrates quite clearly that swaps involving "Administrator Experience" are high-risk, low-distortion swaps. Similarly, the bottom left panel in Figure 7 identifies swaps involving "Race" as having low risk but high distortion, while the upper right panel shows that swaps involving "Sex" are moderate with respect to both risk and distortion. Collectively, these three cases comprise most of the risk–distortion frontier, and so "Administrator Experience", "Race" and "Sex" are plausible candidates for a single-attribute swap.

School Attributes

Attribute Name	Categories
Enrollment	0–250, 250–500, 500–1000, 1000–5000
FTE Classroom Teachers	0–200, 200–400, 400–600, 600–1500
Locale	Central city, Mid-size central city, Urban fringe of large city, Urban fringe of mid-size city, Large town, Small town, Rural
Region	Northeast, Midwest, South, West

Administrator Attributes

Attribute Name	Categories
Years Experience	0–2, 2–5, 5–35
Annual Salary	$0–50,000, $50,000–75,000, $75,000–120,000
Sex	M, F
Race	Non-white, White

Table 3: Attributes and attributes categories for the NCES data.

The bottom left panel in Figure 7, by contrast, shows a poor choice of swap attribute—"School Enrollment:" swaps involving it are characterized by both high risk *and* high distortion.

6 Discussion

The risk–utility formulation for data swapping is a powerful device for informed selection of an actual swapped data release corresponding to a particular choice of swap rate, swap attributes and constraints. Moreover, use of risk–utility frontiers reduces significantly the scale of the associated decision problems.

A number of issues remain unaddressed, however. One of the most important is how to incorporate domain knowledge in a principled manner into disclosure risk and data utility measurements, or into the overall risk–utility formulation of data swapping. For example, what measures of data distortion can incorporate the domain knowledge that it is more important to avoid distorting one attribute in the database than another? How can disclosure risk measures such as (7) reflect the domain knowledge of how easy it is to link $\mathcal{D}_{\text{post}}(R)$ and $\mathcal{D}_{\text{ext}}$?

A second issue is to broaden the decision problem to allow data swapping to be used in conjunction with other strategies for statistical disclosure limitation. For example, can data swapping and category aggregation be used in conjunction in a way that is superior to either alone, and if so, how? NISS is initiating research on this issue.

Acknowledgements

The research reported here was supported in part by NSF grants EIA–9876619 and IIS–0131884 to the National Institute of Statistical Sciences, and by the National Center for Education Statistics. We thank Adrian Dobra, William "Jimmy" Fulp and Chunhua "Charlie" Liu, the latter two of whom played major roles in development of the Data Swapping Toolkit, for their comments.

References

H. Akaike. Information theory and the extension of the maximum likelihood principle. In B.N Petrov and B.F. Csaki, editors, *Second International Symposium on Information Theory*, pages 267–281. Academiai Kiado, Budapest, 1973.

J. L. Bentley, H. T. Kung, M. Schkolnick, and C. D. Thompson. On the average number of maxima in a set of vectors. *J. ACM*, 25:536–543, 1978.

Y. M. M. Bishop, S. E. Fienberg, and P. W. Holland. *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA, 1975.

M. Boyd and P. Vickers. Record swapping—a possible disclosure control approach for the 2001 UK Census. *Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*, 1999.

Census Bureau. Current Population Survey, 2002. Information available on-line at www.bls.census.gov/cps/cpsmain.htm.

L. H. Cox. Confidentiality problems in microdata release. *Proceedings of the Third Annual Symposium on Computer Applications in Medical Care, IEEE Computer Society*, pages 397–402, 1979.

L. H. Cox and L. V. Zayatz. An agenda for research in statistical disclosure limitation, 1993. Available on-line at www.census.gov/srd/papers/pdf/lvz9301.pdf.

N. A. C. Cressie and T. R. C. Read. Cressie–Read statistic. In S. Kotz and N. L. Johnson, editors, *Encyclopedia of Statistical Sciences, Supplementary Volume*. Wiley, New York, 1988.

T. Dalenius and S. P. Reiss. Data swapping: A technique for disclosure control. *J. Statist. Planning Inf.*, 6:73–85, 1982.

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

A. Dobra, A. F. Karr, and A. P. Sanil. Preserving confidentiality of high-dimensional tabular data: Statistical and computational issues. *Statist. and Computing*, 13(4):363–370, 2003.

J. Domingo-Ferrer, J. M. Mateo-Sanz, and V. Torra. Comparing SDC methods for microdata on the basis of information loss and disclosure risk. *presented at UNECE Workshop on Statistical Data Editing*, May, 2001.

G. T. Duncan, S. E. Fienberg, R. Krishnan, R. Padman, and S. F. Roehrig. Disclosure limitation methods and information loss for tabular data. In P. Doyle, J. I. Lane, J. J. M. Theeuwes, and L. V. Zayatz, editors, *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, pages 135–166. Elsevier, Amsterdam, 2001.

G. T. Duncan, S. A. Keller-McNulty, and S. L. Stokes. Disclosure risk vs. data utility: The R-U confidentiality map. *Management Sci.*, 2004. Submitted for publication.

D. E. Edwards and T. Havraneek. A fast procedure for model search in multidimensional contingency tables. *Biometrika*, 72:339–351, 1985.

S. E. Fienberg, U. E. Makov, and R. J. Steele. Disclosure limitation using perturbation and related methods for categorical data. *J. Official Statist.*, 14:485–511, 1998. With discussion.

S. E. Fienberg and J. McIntyre. Data swapping: Variations on a theme by Dalenius and Reiss. In J. Domingo-Ferrer and V. Torra, editors, *Privacy in Statistical Databases '2004*. Springer–Verlag, New York, 2004.

S. Gomatam and A. F. Karr. Distortion measures for categorical data swapping. Technical Report 131, National Institute of Statistical Sciences, 2003. Available on-line from www.niss.org/downloadabletechreports.html.

S. Gomatam, A. F. Karr, J. P. Reiter, and A. P. Sanil. Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access analysis servers. *Statist. Sci.*, 2004. To appear. Available on-line at www.niss.org/dgii/technicalreports.html.

A. F. Karr, A. Dobra, and A. P. Sanil. Table servers protect confidentiality in tabular data releases. *Comm. ACM*, 46(1):57–58, 2003.

J. J. Kim and W. Winkler. Masking microdata files. *Proceedings of the Survey Research Methods Section, American Statistical Association*, pages 114–119, 1995.

H. T. Kung, F. Luccio, and F. P. Preparata. On finding the maxima of a set of vectors. *J. ACM*, 22: 469–476, 75.

D. Lambert. Measures of disclosure risk and harm. *J. Official Statist.*, 9:313–331, 1993.

L. Le Cam and G. L. Yang. *Asymptotics in Statistics*. Springer–Verlag, New York, 1990.

National Institute of Statistical Sciences. Data Swapping Toolkit, 2003a. Available on-line at www.niss.org/software/dstk.html.

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

S. P. Reiss. Practical data-swapping: The first steps. *ACM Trans. Database Systems*, 9(1):20–37, 1984.

A. P. Sanil, S. Gomatam, A. F. Karr, and C. Liu. NISSWebSwap: A Web Service for data swapping. *J. Statist. Software*, 8(7), 2003.

Gideon Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6:461–464, 1978.

N. L. Spruill. Measure of confidentiality. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pages 260–265, 1982.

A. Takemura. Local recoding and record swapping by maximum weight matching for disclosure control of microdata sets. *J. Official Statist.*, 18:275Ű–289, 2002.

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

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

J. Whittaker. *Graphical Models in Applied Multivariate Statistics*. Wiley, New York, 1990.

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

L. C. R. J. Willenborg and T. de Waal. *Elements of Statistical Disclosure Limitation*. Springer–Verlag, New York, 2000.

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

W. E. Winkler. Re-identification methods for evaluating the confidentiality of analytically valid microdata. *Res. Official Statist.*, 1:87–104, 1998.

W. E. Yancey, W. E. Winkler, and R. H. Creecy. Disclosure risk assessment in perturbative microdata. *Inf. Control in Statist. Databases*, 2002.

A. M. Zaslavsky and N. J. Horton. Balancing disclosure risk against the loss of nonpublication. *J. Official Statist.*, 14(4):411–419, 1998.

A The Swapping Algorithm

Let n be the number of records in the pre-swap database $\mathcal{D}_{\text{pre}}$. Then the Data Swapping Toolkit swapping algorithm operates in the following manner:

1. Initially, mark all records as unswapped and set $\mathcal{R}$, the number of swapped records, to zero.
2. The r be the user-specified swap rate, and let RTS $= \lfloor r \times n \rfloor$ be the number of records to be swapped.
3. Select a record R_1 at random from the current set of unswapped and "not unswappable" records.
4. Select a second record R_2 at random from the current set of *all* unmarked (as either swapped or unswappable) records.
5. Determine whether the swap is a *true* swap: R_1 and R_2 must differ on both the swap attribute and as least one unswapped attribute. If not, return to Step 4.
6. Determine whether equality and inequality swapping constraints are satisfied. If not, return to Step 4.
7. If no feasible candidate R_2 can be found, go to Step 10.
8. Otherwise, interchange the swapped attribute(s) between R_1 and R_2, mark *both* as swapped, and set $\mathcal{R} = \mathcal{R} + 2$.
9. If $\mathcal{R} <$ RTS, return to Step 3. Otherwise, the swapping is complete. For Batch Swaps, label the swap a "success."
10. Mark R_1 as unswappable (no other unswapped record in the database can be swapped with it). If any records remain that are both unswapped and "not unswappable," return to Step 3. Otherwise, terminate the algorithm and label the swap a "failure."

Note that this algorithm does not take into account any "weights" Takemura (2002) in selection of swap pairs. The risk-utility formulation is this paper extends immediately, as long as the method of calculating weights is not a decision variable.

B Two-Dimensional Measures of Data Distortion

In some cases, attention might focus on two-way relationships between attributes. Two approaches are possible. The first simply restricts the Hellinger distance, total variation distance and entropy change to two-dimensional marginals of $\mathcal{D}_{\text{pre}}$ and $\mathcal{D}_{\text{post}}(R)$. Alternatively, measures of distortion specific to two attributes, which are sometimes termed measures of association, can be employed. Two such measures have been investigated.

Cramer's V, which is based on the χ^2 statistic for a $m \times n$ contingency table:

$$V = \sqrt{\frac{\chi^2}{N \min(m-1, n-1)}},$$

where χ^2 is the usual χ^2 statistic to test for independence. Its values lie between 0 and 1—a value of 0 indicates no association, whereas a value of 1 indicates perfect association. It is more difficult to interpret values between the extremes. Cramer's V has been used in the context of data swapping to assess the effects of swapping within geographically defined subsets of the population (Boyd and Vickers, 1999). To measure data distortion, one can employ

$$\mathrm{CV}_{ij}(\mathcal{D}_{\mathrm{pre}}, \mathcal{D}_{\mathrm{post}}(R)) = V_{ij}(\mathcal{D}_{\mathrm{pre}}) - V_{ij}(\mathcal{D}_{\mathrm{post}}(R)), \tag{15}$$

where i and j represent attributes. Positive values of $\mathrm{CV}_{ij}(\mathcal{D}_{\mathrm{pre}}, \mathcal{D}_{\mathrm{post}}(R))$ indicate that swapping has weakened the association between attributes i and j.

Contingency coefficients. Pearson's contingency coefficient C also measures association:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}},$$

where χ^2 is again the usual χ^2 defined for the test of independence. Values of C lie between 0 and 1, but the upper limit depends on m and n, so it is difficult to compare tables of different sizes. Like Cramer's V, C also suffers from the difficulty of interpretation for intermediate values. We then define

$$\mathrm{CC}_{ij}(\mathcal{D}_{\mathrm{pre}}, \mathcal{D}_{\mathrm{post}}(R)) = C_{ij}(\mathcal{D}_{\mathrm{pre}}) - C_{ij}(\mathcal{D}_{\mathrm{post}}(R)). \tag{16}$$

Positive values of $\mathrm{CC}_{ij}(\mathcal{D}_{\mathrm{pre}}, \mathcal{D}_{\mathrm{post}}(R))$ indicate that swapping has weakened the association between i and j.

Tables containing the numerical values of these five distortions for the CPS-8D data appear in Gomatam and Karr (2003). Overall there is significant consistency in the conclusions drawn from the different measures. *Age* and *Income* are the preponderant maximizers for both Hellinger distance and entropy change; *MS* also plays a significant role for total variation distance. *Race* and *Edu* are the primary minimizers for all three of these measures. The behavior of CV and CC is primarily like that of 2-way Hellinger distance: *Income* is most likely to be a maximizer and *Edu* is most likely to be a minimizer. However, *MS* plays a stronger role than *Age* in maximizing CV and CC.

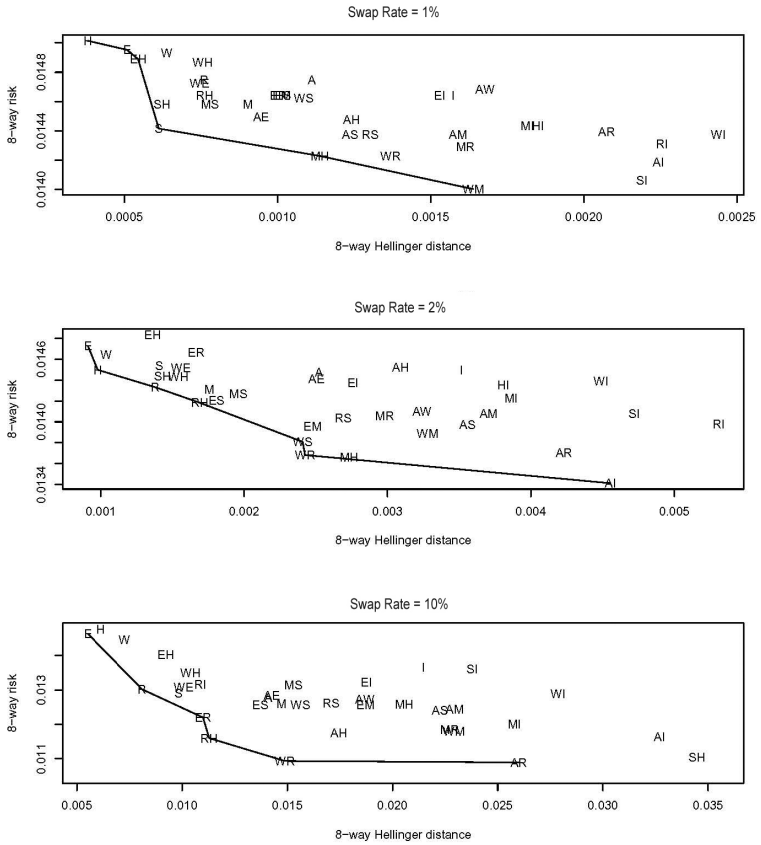

Figure 3: Risk–distortion scatterplots for 108 candidate releases from the CPS-8D database. Swap attributes for each case are identified using the abbreviations in Table 1. *Top:* swap rate = 1%. *Middle:* swap rate = 2%. *Bottom:* swap rate = 10%. Each scatterplot contains 36 candidate releases representing all choices of one or two swap attributes.

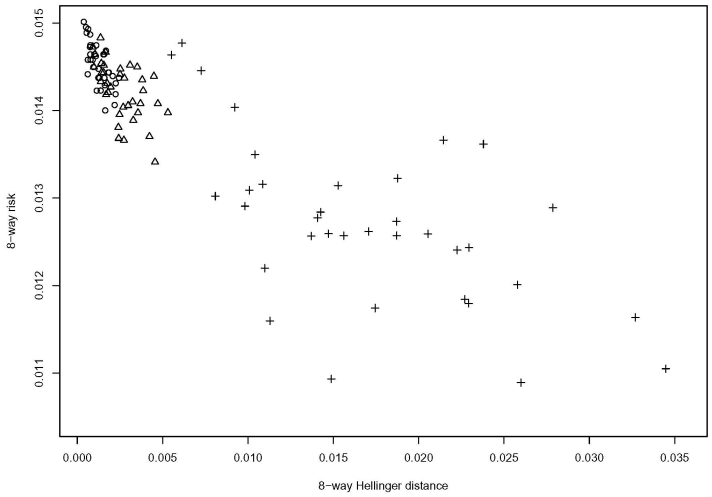

Figure 4: Risk–distortion scatterplots for 108 candidate releases from the CPS-8D database. Three swap rates (1%–circles, 2%—triangles and 10%—plus signs) are shown, and for each, there are 36 candidate releases representing all choices of one or two swap attributes.

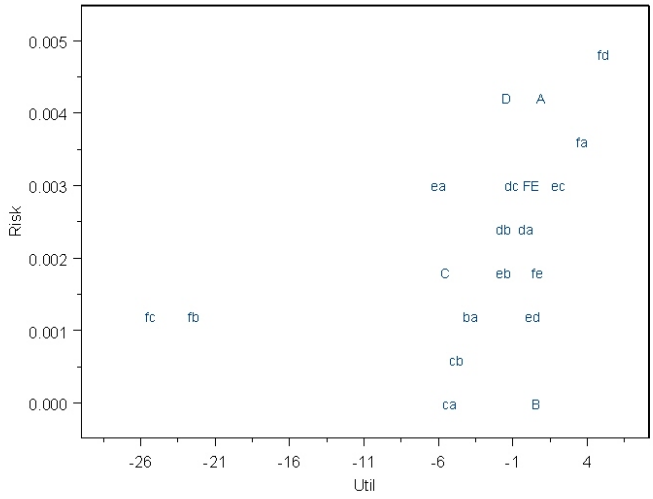

Figure 5: Risk–utility plot for the Czech automobile worker database of Table 2, using the inference-based utility $\mathbf{DU}_{\mathrm{llm}}(R)$ of (12) and the small cell count risk measure in (6). Points are labeled by swap attributes—A, . . ., F for single-attribute swaps and fe, . . ., ba for two-attribute swaps.

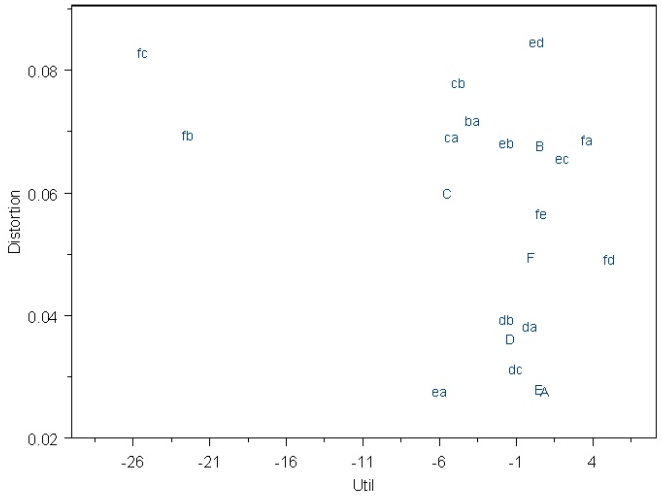

Figure 6: Relationship between inference-based utility $\mathbf{DU}_{\text{llm}}(R)$ from (12) and Hellinger distance-based data distortion from (9) for the Czech automobile worker database of Table 2. Points are labeled as in Figure 5

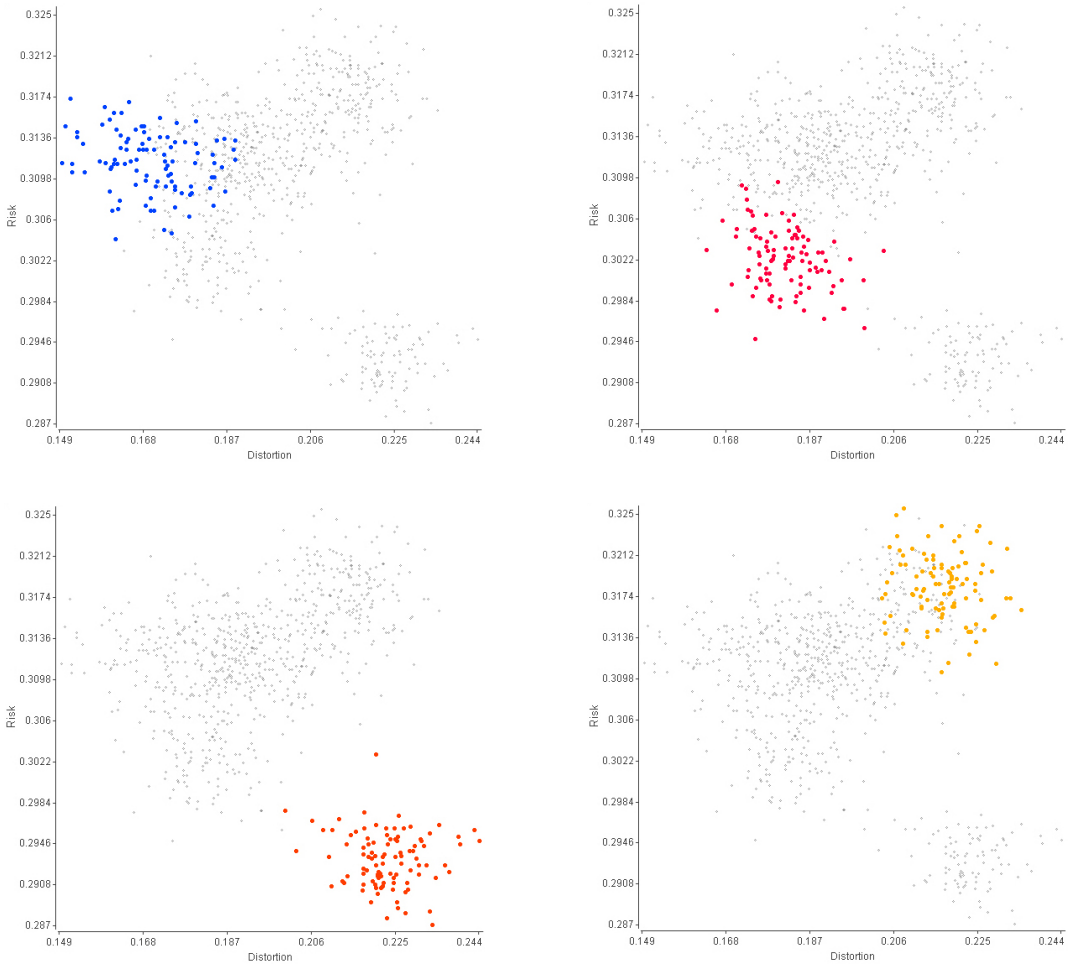

Figure 7: Scatterplot of (distortion, risk) values for 800 swaps of the NCES data. *Upper left:* swaps involving "Administrator Experience" highlighted. *Upper right:* swaps involving "Sex" highlighted. *Bottom left:* swaps involving "Race" highlighted. *Bottom right:* swaps involving "School Enrollment" highlighted.