

**A Risk-Utility Framework
for Categorical Data Swapping**

Shanti Gomatam, Alan F. Karr and Ashish Sanil^a

National Institute of Statistical Sciences

{sgomatam,karr,ashish}@niss.org

March 13, 2003

^aSupport for this research was provided by National Science Foundation grant EIA-9876619 to NISS, and by the National Center for Education Statistics.

OUTLINE

- What is data swapping (DS)?
- Effect of DS
- Data releases; Risk and Utility
- RU frontiers and optimal releases
- Illustrative example
- Conclusions

Background

- Data are collected on subjects and transferred to the disseminator/agency, under conditions of protecting confidentiality.
- Disseminator then makes the data available to users, making sure that the data are protected in such a way that intruders cannot compromise the privacy of data subjects.
- As data disseminators wish to provide as much information as possible to the data user while satisfying the mandate of confidentiality.

Data Swapping

Conceptualize the microdata/data file as a matrix, with the rows (records) representing subjects, and columns representing variables (attributes)

Data swapping involves “switching column values for pairs of rows” for a selected fraction of rows.

NISS has created software available as a web service (see links at <http://www.niss.org>) to implement swapping.

Simple example:

Rec. No.	AvgHrs	EmpTyp	Sex	MarStat
1	< 40	Gov	M	M
2	40	SelfEmp	F	UM
3	< 40	Priv	F	M
4	> 40	Priv	M	M

Rec. No.	AvgHrs	EmpTyp	Sex	MarStat
1	< 40	Gov	M	M
2	> 40	SelfEmp	F	UM
3	< 40	Priv	F	M
4	40	Priv	M	M

Some terms:

swap variables: Subset of variables that will be swapped.

swap rate: The (half-)fraction (usually small) of the total records in the microdata that are swapped.

constraining variables: Variables whose values define the feasibility of a swap pair.

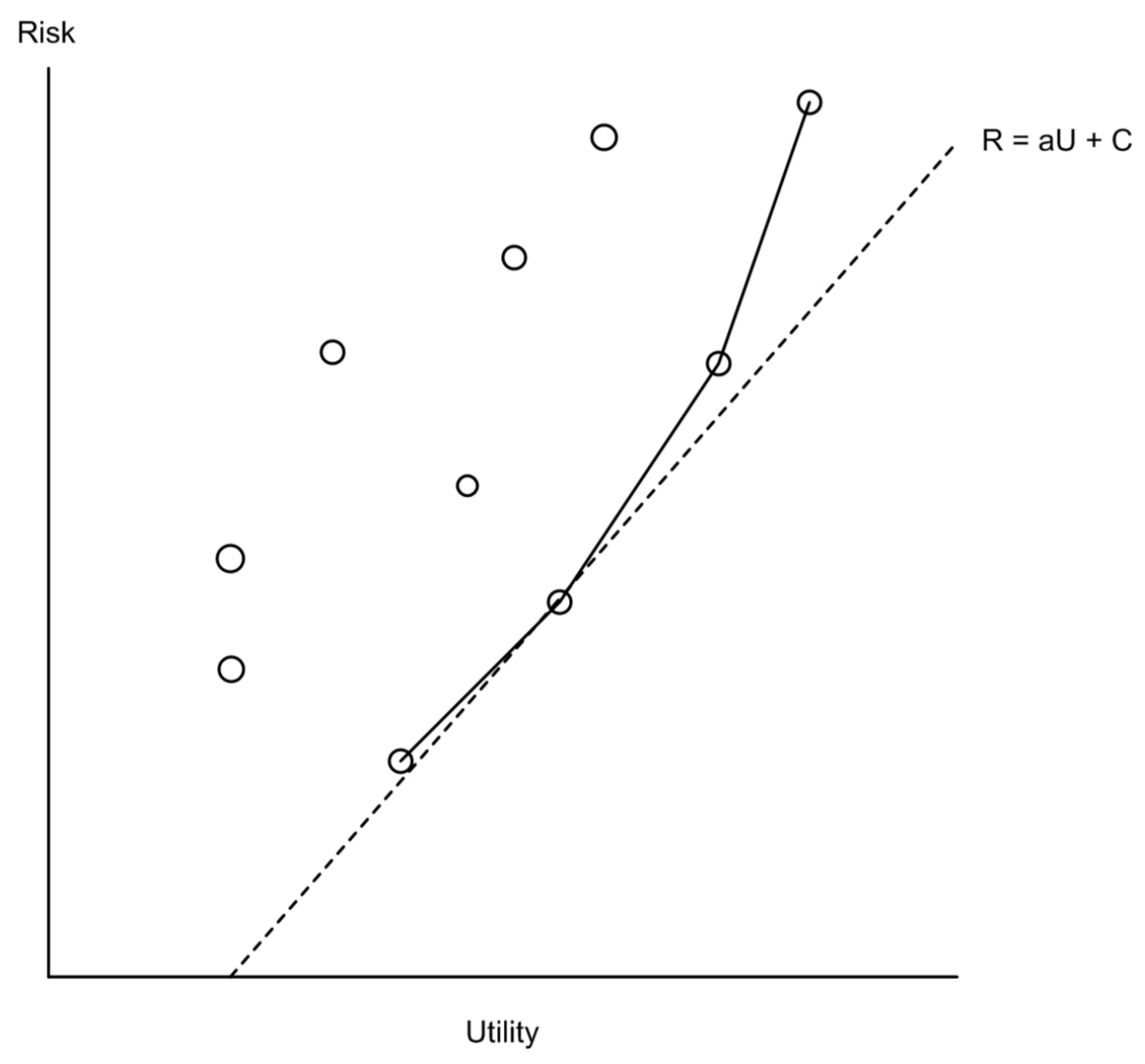
Constrained swaps are swaps that contain constraining variables, when there are no constraining variables involved we call the swapping process “unconstrained.”

Effect of Data Swapping

- Univariate marginal distributions are preserved.
- Joint distributions not involving swap variables or only involving swap variables are preserved.
- Joint distributions that involve both swap and non-swap variables may be distorted.

Selecting a Release

- Need to pick one of several candidate data releases.
- Possible parameters: swap variables, swap rate, different realizations of the swapping (which involves randomization), ...
- Each release has a *risk* of compromising confidentiality and a data *utility* to users.
- Choice of release take place within a *risk-utility* framework, balance risk and utility to pick optimal release.



Risk and Utility measures

Risk measure: Proportion of *unswapped* records in small count cells in the table created from post-swap data:

$$\text{Risk} = \frac{\sum_{C_1, C_2} \text{Number of unswapped records}}{\text{Total number of unswapped records}},$$

where C_1 and C_2 are the cells in the full data table with counts of 1 and 2.

Utility measure:

$$H(f, g) = \sqrt{\frac{1}{2} \sum_C \left(\sqrt{f(C)} - \sqrt{g(C)} \right)^2},$$

where the sum is over cells C in the full data table.

Data Illustration for Rate and Swap

Variable choice: Using 48,842 observations from Current Population Survey.

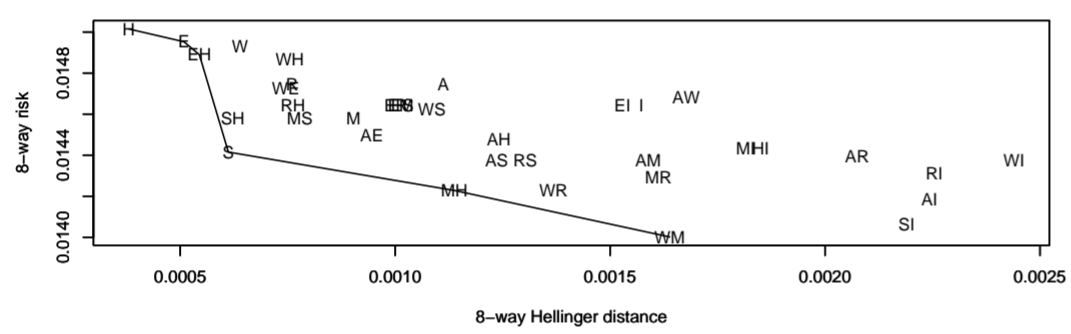
Variable Name	Abbreviation	Categories
Age (in years) (<i>Age</i>)	A	<25, 25–55, >55
Employer Type (<i>WrkTyp</i>)	W	Govt., Priv., Self-Emp., Other
Education (<i>Educ</i>)	E	<HS, HS, Bach, Bach+, Coll
Marital Status (<i>MarStat</i>)	M	Married, Other
Race (<i>Race</i>)	R	White, Non-White
Sex (<i>Sex</i>)	S	Male, Female
Average Weekly Hours Worked (<i>Hours</i>)	H	< 40, 40, > 40
Annual Salary (<i>Income</i>)	I	<\$50K, \$50K+

Unconstrained Swaps: For CPS-8d data with 48,842 observations, tried:

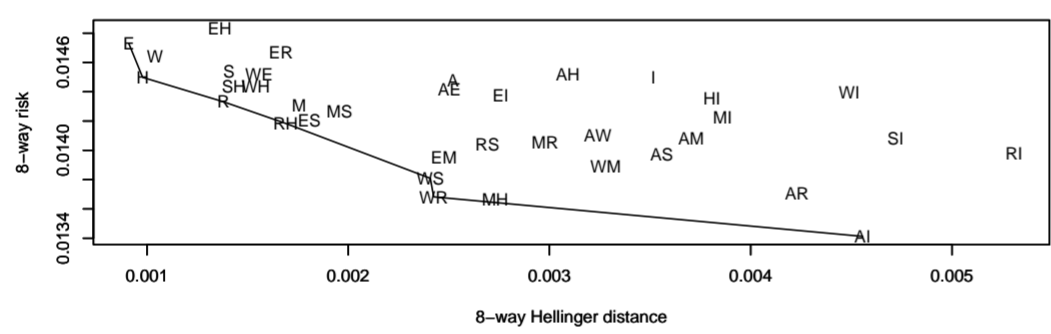
- all single and two-variable swaps
- 3 swap rates (0.005, 0.01, 0.05)

Graphs show frontiers for each rate separately, and joint frontier for all three rates.

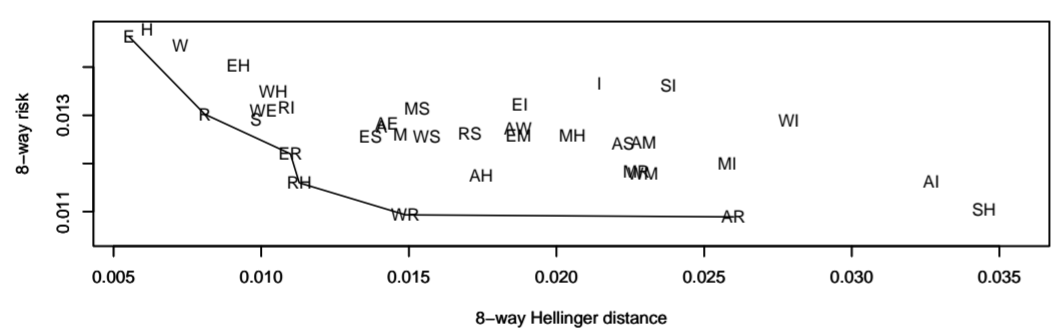
(a) Swap rate= 0.005



(b) Swap rate= 0.01



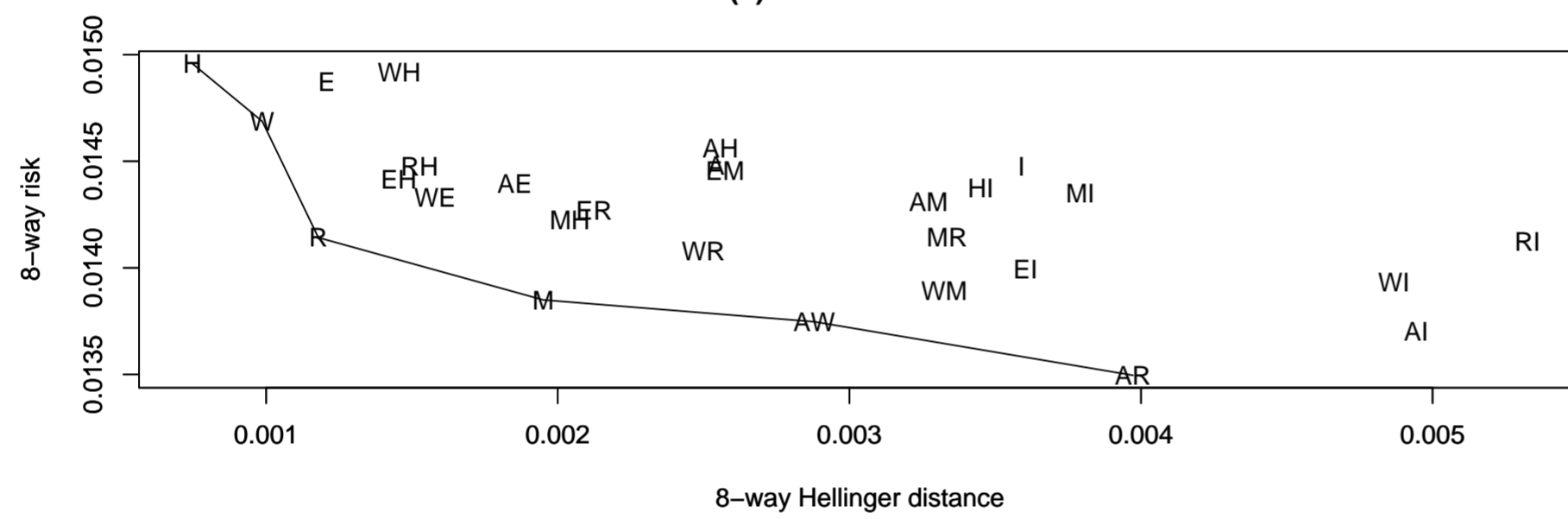
(c) Swap rate= 0.05



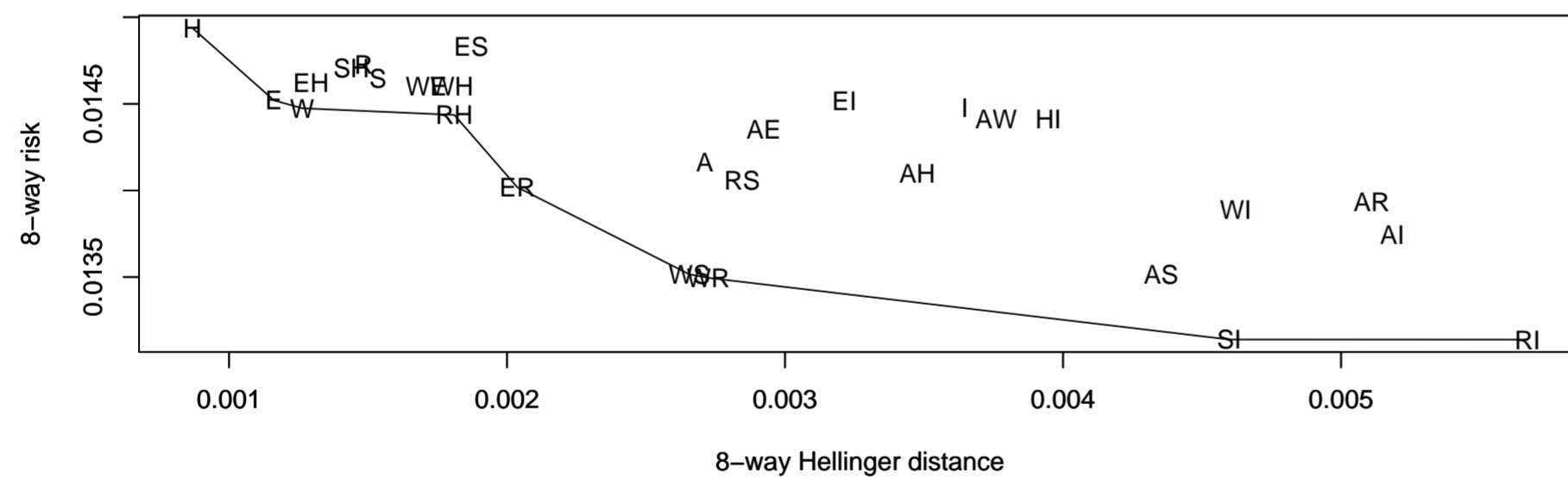
Constrained Swaps: For CPS-8d data and swap rate=0.01, tried two different constrained swaps:

- (a) Sex constrained to be the same for any feasible swap
- (b) Marital Status constrained to be different for any feasible swap

(a) S constrained.



(b) M constrained.



Concluding points

- Framework to discriminate among post-swap data releases.
- Illustrated for candidate releases corresponding different choices of swap variable(s) and swap rate, but it applies equally well to other parameters.
- Can even be used to compare multiple disclosure methods for which the same risk and utility measures are reasonable.

References, contact info., etc.

- This and related work available as
Technical Reports No. 126, 131, 132, 134 at
NISS
(<http://www.niss.org>)
- NISSWebSwap at
<http://www.niss.org/WebServices/dg/WebSwap.html>
- Contact info.
sgomatam@niss.org