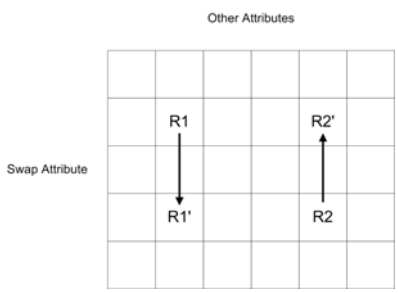
NISS Data Swapping Toolkit (DSTK)

Alan Karr
March 3, 2005

Data Swapping

- Technique for statistical disclosure limitation (SDL), applied at microdata level
- Basic idea: switch subset of attributes between randomly selected pairs of records
- Used by: Census, …
- Positive side: reduces disclosure risk
  - Intruder cannot be certain that any record is real
- Negative side: distorts data
  - Reduces utility

Tabular View



Example Swap for CPS-8

Record	Age	EmplType	Educ	MarStat	Race	Sex	AveHours	Salary
1	<25	Gov	HS	Marr	W	M	40	<$50K
2	25–55	SE	Bach	Marr	NW	M	>40	<$50K
3	25–55	Gov	Bach+	Unmarr	NW	F	>40	>$50K
4	>55	Priv	Bach	Unmarr	W	F	>40	<$50K
5	<25	Other	SomeColl	Marr	W	M	40	>$50K
6	>55	Priv	Bach+	Marr	NW	F	40	>$50K

Record	Age	EmplType	Educ	MarStat	Race	Sex	AveHours	Salary
1	>55	Gov	HS	Marr	W	M	40	<$50K
2	25–55	SE	Bach	Marr	NW	M	>40	<$50K
3	<25	Gov	Bach+	Unmarr	NW	F	>40	>$50K
4	>55	Priv	Bach	Unmarr	W	F	>40	<$50K
5	25-55	Other	SomeColl	Marr	W	M	40	>$50K
6	<25	Priv	Bach+	Marr	NW	F	40	>$50K

Technical Aspects

- Parameters
  - Swap rate: typical value = 5%
  - Swap attribute(s)
  - Optionally, constraints on unswapped attributes
- Distortion effects
  - No change to joint distribution of swap attributes
  - No change to joint distribution of unswapped attributes
  - Change to joint distributions that involve both swap and unswapped attributes

Risk-Utility Formulation: Generalities

- Components
  - Database $\mathcal{D}$
  - Set $\mathcal{R}$ of *candidate releases* R = f($\mathcal{D}$)
  - Disclosure risk function **DR**(R)
  - Data utility function **DU**(R)
- Goal: Select the "best release"

Selection Procedures

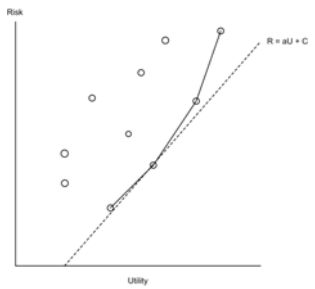
- Maximize utility subject to upper bound on risk

$$R^* = \arg\max_{R \in \mathcal{R}} \mathbf{DU}(R)$$
$$\text{s.t. } \mathbf{DR}(R) \leq \alpha.$$

- Select from *risk-utility frontier* defined by the partial order

$$R_1 \preceq_{\text{RU}} R_2 \quad \Leftrightarrow \quad \mathbf{DR}(R_2) \leq \mathbf{DR}(R_1) \text{ and } \mathbf{DU}(R_2) \geq \mathbf{DU}(R_1)$$

Conceptual Risk-Utility Frontier



Risk-Utility Formulation for Data Swapping

- Release R defined by
  - Swap attributes
  - Swap rate
  - Constraints on unswapped attributes
- Disclosure risk measure

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

- Utility measure

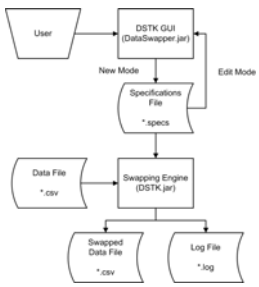
$$\mathbf{DU}(R) = -\mathbf{DD}(R) = -\text{HD}(\mathcal{D}_{\text{pre}}, \mathcal{D}_{\text{post}}(R)),$$

DSTK Basics

- Written in Java by
  - Ashish Sanil
  - Jimmy Fulp
  - Charlie Liu
- Available at www.niss.org/software/dstk.html
- Three components
  - GUI for single swaps
  - Batch swap package
  - [Integrated Batch Swapper]
  - Frontier visualizer

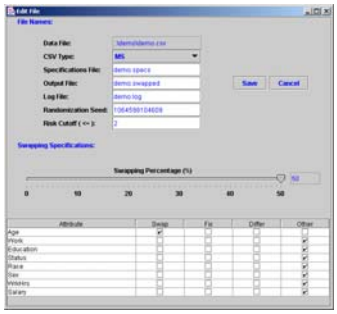
The Single Swap GUI

- Select
  - CSV data file
  - CSV output file
  - Log file
  - Swap specifications
  - Risk cutoff
  - …
- Execute swap



Single Swap GUI—2

Project Editor



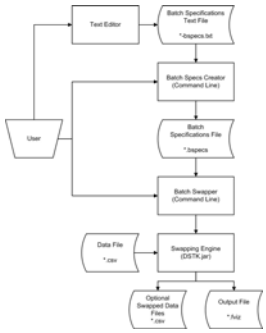
The Specifications File

```
#Fri Sep 26 13:41:46 EDT 2003
num.front=1
data.type=1
spec.file=demo.specs
attribute.specs=S,0,0,0,0,0,0,0
output.file=demo.swapped
csv.type=MS
log.file=demo.log
data.file=..\\demo\\demo.csv
random.seed=1064598104609
swap.percentage=50.0
risk.cutoff=2
```

Doing the Swap

The Batch Swapper

- Batch specifications creator
- Batch swapper
- Integrated batch swapper (IBS)



Batch Specifications Creator

- Command line invocation
- Input = batch specifications text file *-bspecs.txt containing
  - File names
  - Swap specifications
    - Release parameters
    - Experiment parameters
  - Risk cutoff
  - …
- Output = batch specifications file *.bspecs containing input for batch swapper

The Batch Specifications Text File

```
data.file=demo.csv
swap.rates=0.01,0.02
#swap.options=oneway
#swap.options=twoway
swap.options=oneway,twoway
specs.file=demo.bspecs
output.file=demo.fviz
#save.dir=
#csv.type=ISO
#risk.cutoff=
record.id=true
#weight=true
#weight.category=true
```

The Batch Specifications File

```
#demo.bspecs was created from demo-bspecs.txt
#Thu Oct 02 13:26:04 EDT 2003
record.id=true
weight.category=false
output.file=demo.fviz
weight=false
csv.type=MS
swap.options=oneway,twoway
data.file=demo.csv
specs.file=demo.bspecs
swap.rates=0.01,0.02
risk.cutoff=2
!0.01,S,0,0,0,0,0,0,0
!0.01,0,S,0,0,0,0,0,0
!0.01,0,0,S,0,0,0,0,0
!0.01,0,0,0,S,0,0,0,0
!0.01,0,0,0,0,S,0,0,0
```

The Batch Swapper

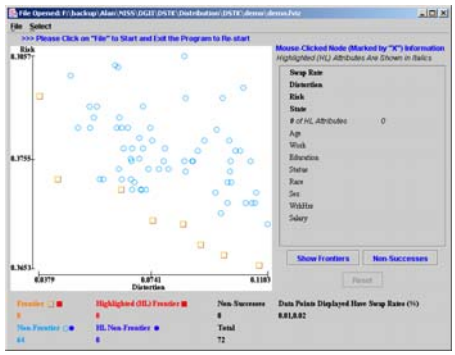
- Input: batch specifications file *.bspecs
- Output
  - Swapped data files
  - Summary output file *.fviz usable
    - As input to frontier visualizer
    - Directly

```
Age,Work,Education,Status,Race,Sex,WrkHrs,Salary,Rate,Dist,Risk,Flag,Seed
S,0,0,0,0,0,0,0,0.01,0.06978024120239666,0.37777777777777777,1,1064598464562
0,S,0,0,0,0,0,0,0.01,0.0638325786079739,0.3727272727272274,1,1064598464859
0,0,S,0,0,0,0,0,0.01,0.05624438888941473,0.3808080808080808,1,1064598464921
0,0,0,S,0,0,0,0,0.01,0.03794299301838207,0.38181818181818183,1,1064598465000
0,0,0,0,S,0,0,0,0.01,0.0640389737288213,0.3737373737373376,1,1064598465078
```

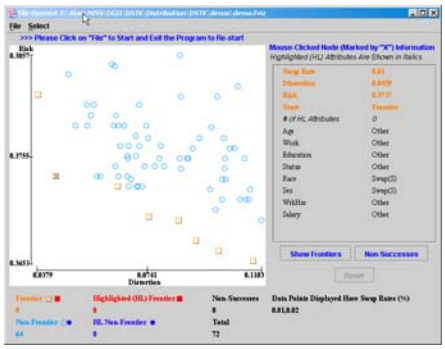
Frontier Visualizer—Functionality

- Scatterplot of (**DD**, **DR**) values
  - Display of individual values
  - Transformation of axes
  - “Show Frontier”
- Selection
  - Rate
  - Attributes
- Save and print

Frontier Visualizer—Main Screen

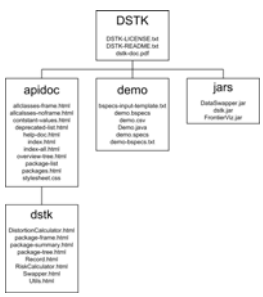


Frontier Visualizer—Drilldown



System Requirements

- Windows (2000, XP) or Linux
- JRE (J2SE) 1.4.1 or higher
  - JRE in PATH
  - CLASSPATH defined
- Data must fit into memory



Thanks to

- BLS, Census, NCES, NCHS, NSF for support
- Shanti Gomatam for algorithm development