

NISS

Data Swapping: A Risk-Utility Framework and Web Service Implementation

Shanti Gomatam, Alan F. Karr,
Chunhua “Charlie” Liu and Ashish P. Sanil
{sgomatam,karr,cliu,ashish@niss.org}

Outline

- Introduction to Data Swapping
- Risk-Utility (Risk-Distortion) Framework
- NISS WebSwap
 - Web services implementation
 - GUI-based, lightweight client
 - Frontier visualization tool
- Future research

Data Swapping

- Technique for statistical disclosure limitation (SDL), applied at microdata level
- Basic idea: switch subset of attributes between randomly selected pairs of records
- Rationale: reduces disclosure risk—intruder cannot be certain that any record is real
- Used by: Census, ...
- Side effect: distorts data, reducing utility

Tabular View

Other Attributes

Swap Attribute

	R1			R2'	
	R1'			R2	

Technical Aspects

- Parameters
 - Swap rate: typical value = 5%
 - Swap attributes
 - 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

Example Data Set: CPS-8

- Excerpt from 1993 CPS
 - 48,842 data records
 - 8 categorical attributes
 - 2880 cells in full table
 - 1695 cells with non-zero counts (not realistic!)
 - 735 at risk data elements
 - 361 in cells with count 1
 - 374 in cells with count 2

Example Swap for CPS-8

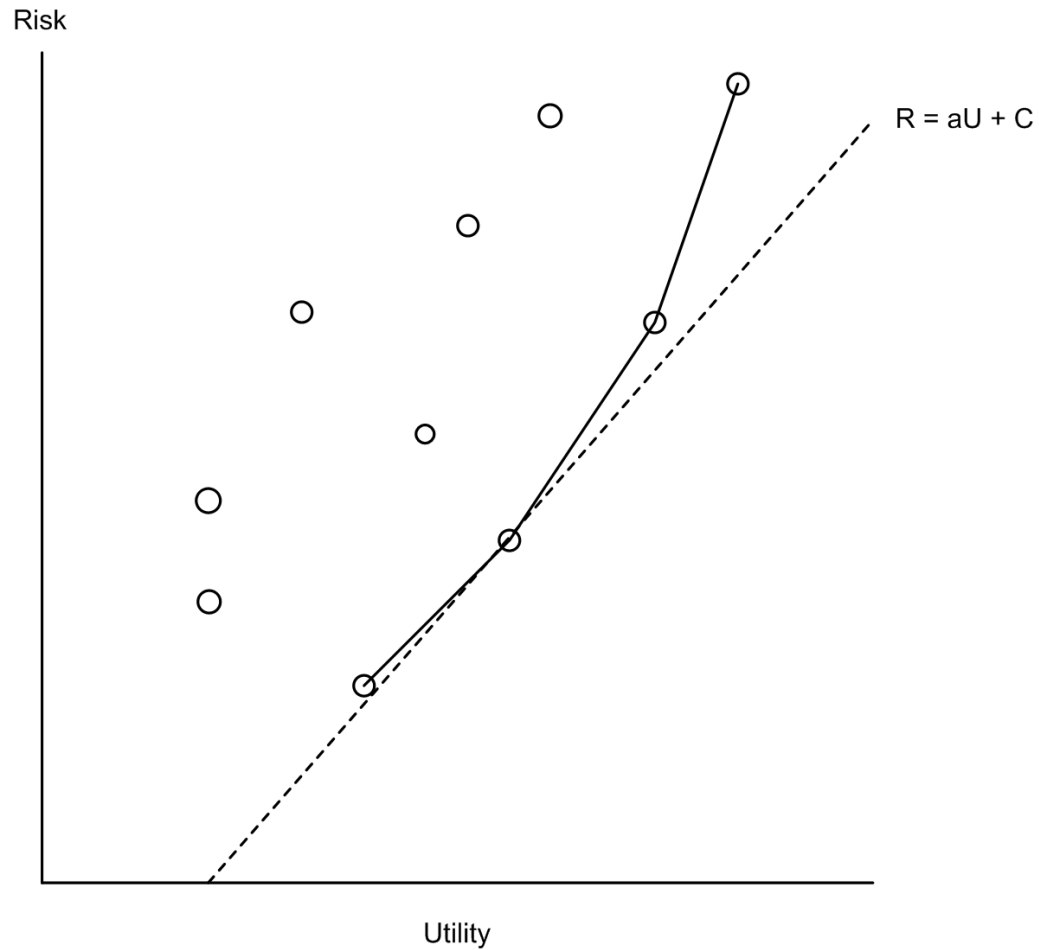
Record	Age	EmplType	Educ	MarStat	Race	Sex	AveHours	Salary
1	<25	Gov	HS	Marr	W	M	40	<\$50K
2	25-55	SE	Bach	Marr	NW	M	>40	<\$50K
3	25-55	Gov	Bach+	Unmarr	NW	F	>40	>\$50K
4	>55	Priv	Bach	Unmarr	W	F	>40	<\$50K
5	<25	Other	SomeColl	Marr	W	M	40	>\$50K
6	>55	Priv	Bach+	Marr	NW	F	40	>\$50K

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

Risk-Utility Framework

- Characterize each candidate release by
 - Disclosure **r**isk
 - Data **U**tility
- Agency wishes to, but cannot simultaneously
 - Minimize risk
 - Maximize utility
- Restrict attention to frontier of undominated releases

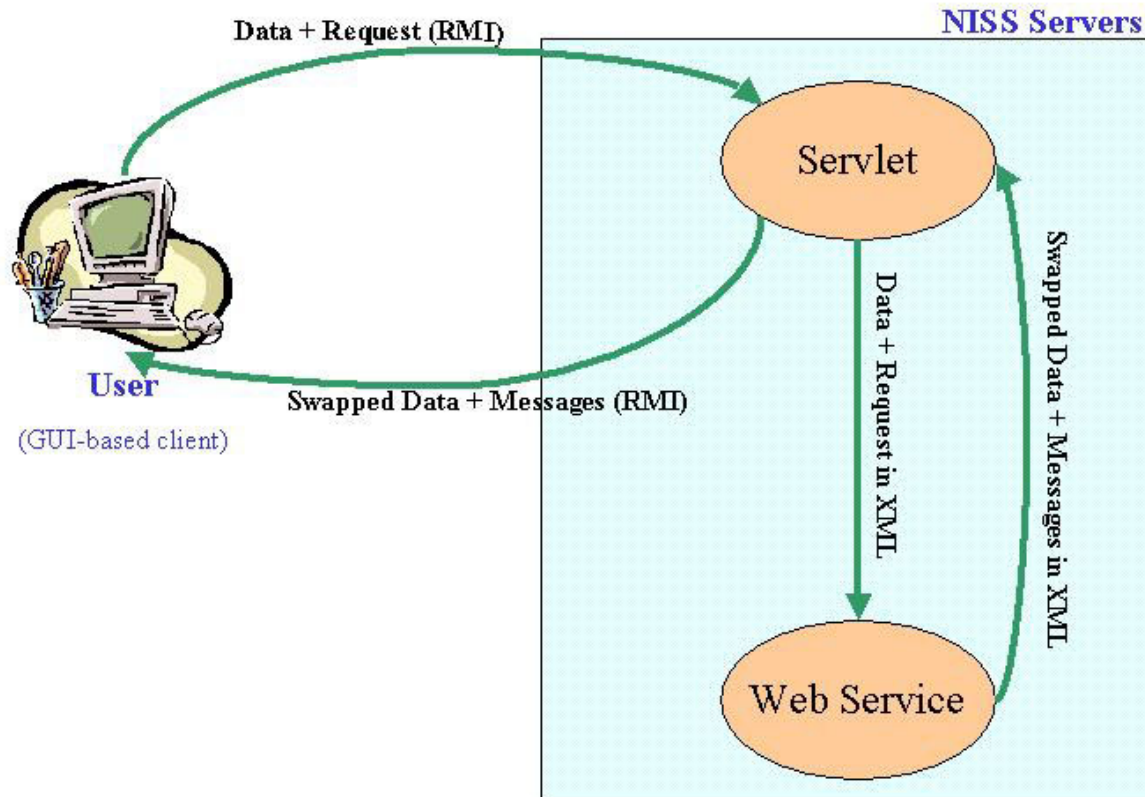
Example Frontier



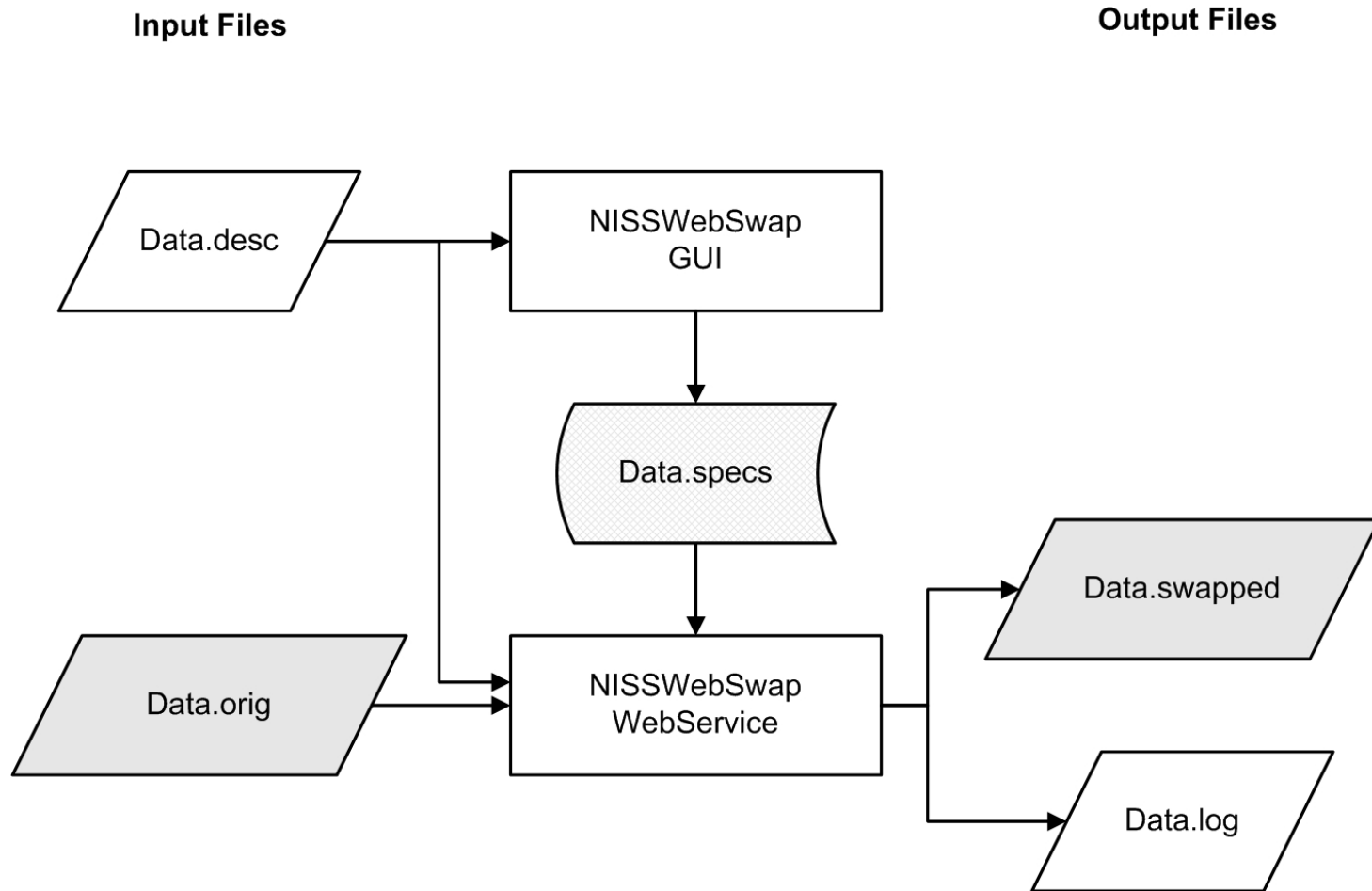
NISS WebSwap Web Service

- Implemented as free Web service, using XML and SOAP to communicate
- WSDL available, supporting discovery, use
- Principal components
 - GUI-based, lightweight client
 - Java application: constructs specification, transmits specification and data file
 - Java servlet on NISS server
 - Server side application: prepares XML request to swapper, returns swapped data and log files to client
 - Web service on NISS server
 - Java application: performs swapping, returns XML to servlet

System Architecture



Logical Architecture



NISS WebSwap WSDL: Excerpt

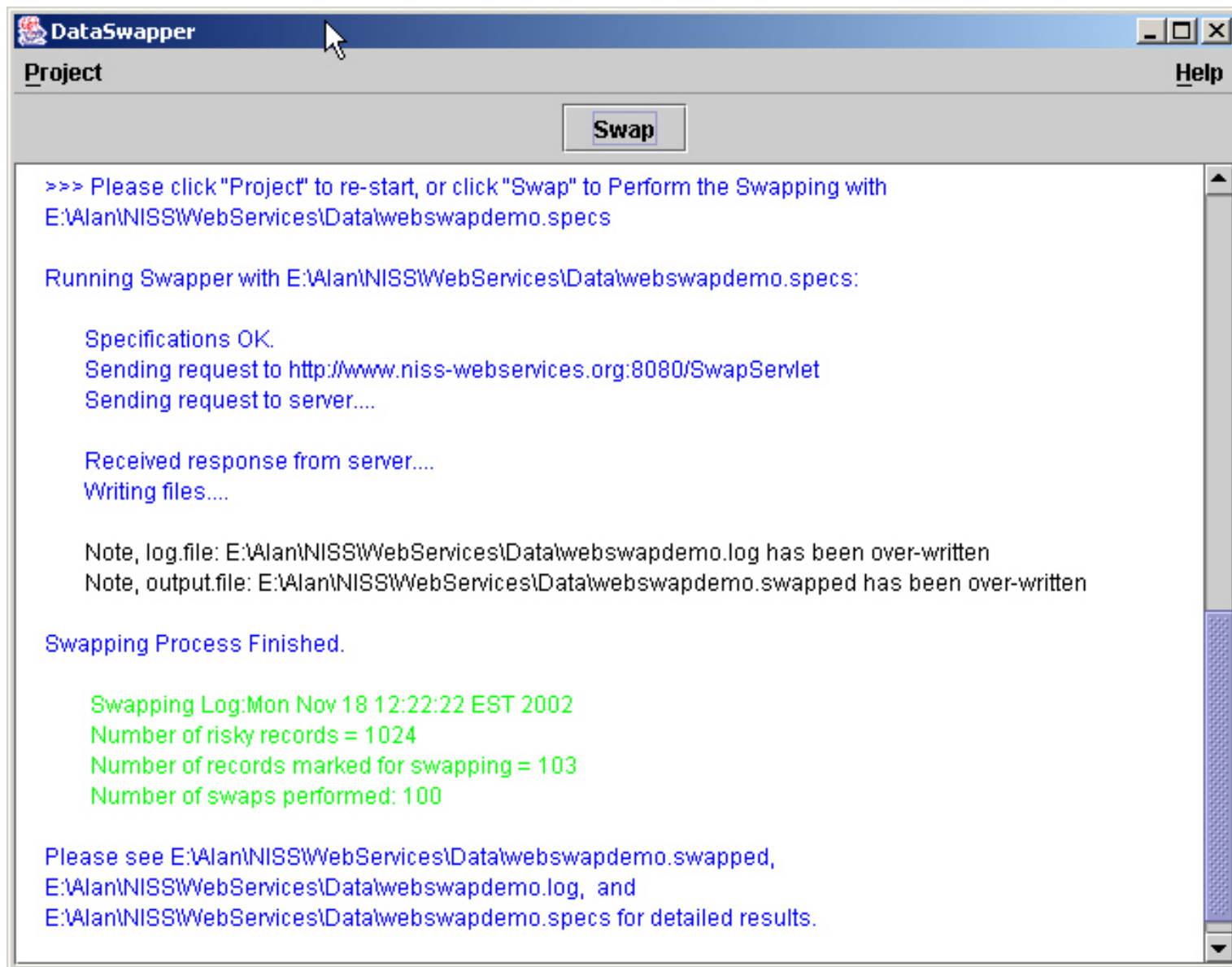
```
<?xml version="1.0" encoding="UTF-8"?>

<definitions name="Swap_dataService"
targetNamespace="http://WebSwap_swap.org/wsdl"
xmlns:tns="http://WebSwap_swap.org/wsdl"
xmlns="http://schemas.xmlsoap.org/wsdl/"
xmlns:soap="http://schemas.xmlsoap.org/wsdl/soap/"
xmlns:ns2="http://WebSwap_swap.org/types"
xmlns:xsd="http://www.w3.org/2001/XMLSchema">
  <types>
    <schema targetNamespace="http://WebSwap_swap.org/types"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xmlns:tns="http://WebSwap_swap.org/types"
      xmlns:soap-enc="http://schemas.xmlsoap.org/soap/encoding/"
      xmlns:wsdl="http://schemas.xmlsoap.org/wsdl/"
      xmlns="http://www.w3.org/2001/XMLSchema">
      <complexType name="SwapData">
        <sequence>
          <element name="numFields" type="int"/>
          <element name="outputFile" type="string"/>
          <element name="numRecords" type="int"/>
          <element name="riskCutoff" type="double"/>
          <element name="data" type="tns:ArrayOfArrayOfstring"/>
          <element name="dataFile" type="string"/>
          <element name="constraints" type="base64Binary"/>
          <element name="log" type="tns:ArrayOfstring"/>
        </sequence>
      </complexType>
    </schema>
  </types>
</definitions>
```

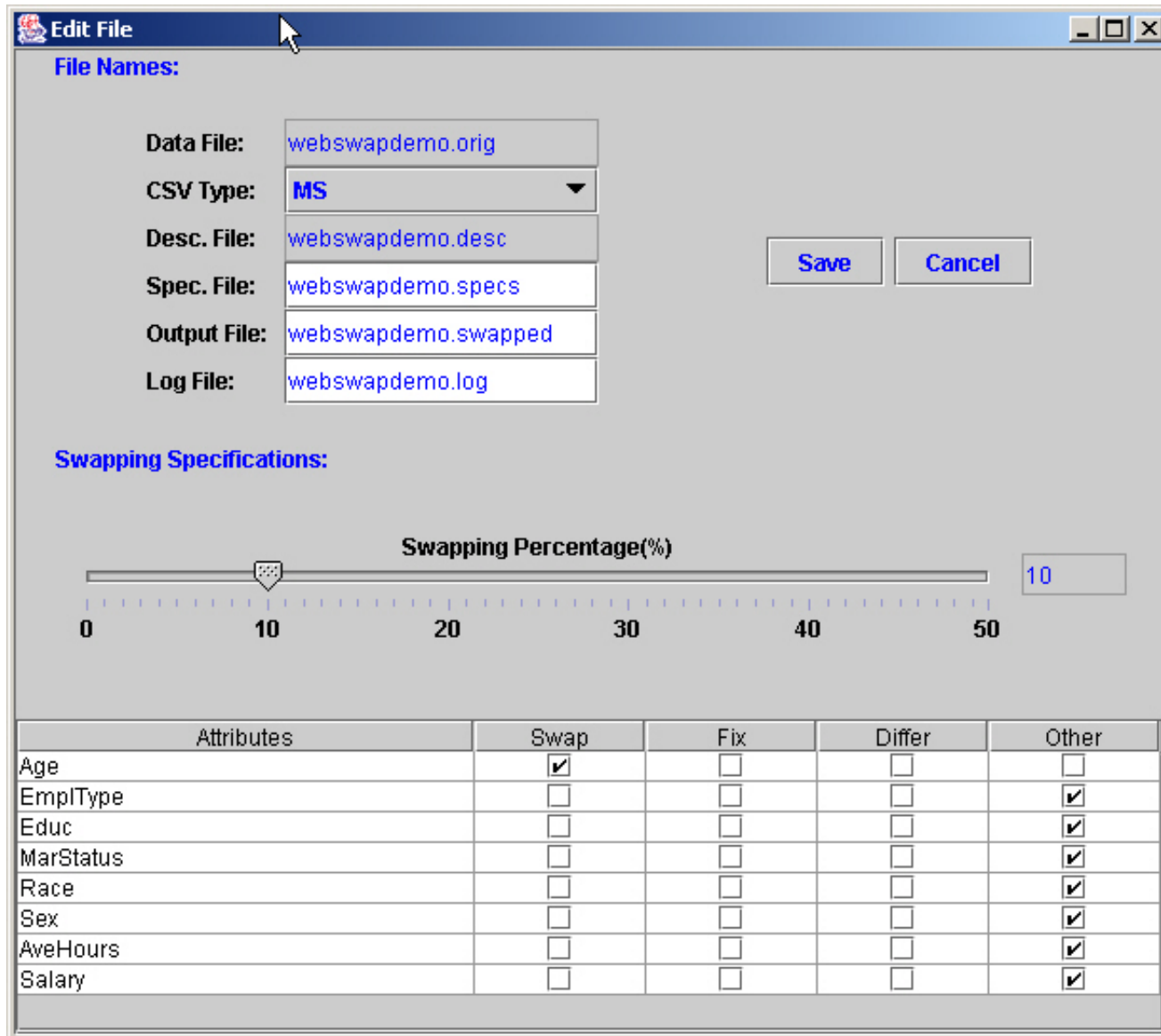
GUI-Based Client

- Inputs
 - CSV data file
 - Description file: metadata
- User constructs swapping specification
 - Swap rate
 - Swap attributes
 - Constraints on unswapped attributes: equality, inequality
- Output: specifications file

GUI Main Window



The Edit Project Window



The screenshot shows a window titled "Edit File" with a mouse cursor pointing to the title bar. The window is divided into two main sections: "File Names:" and "Swapping Specifications:".

File Names:

- Data File: webswapdemo.orig
- CSV Type: MS (dropdown menu)
- Desc. File: webswapdemo.desc
- Spec. File: webswapdemo.specs
- Output File: webswapdemo.swapped
- Log File: webswapdemo.log

Buttons for "Save" and "Cancel" are located to the right of the file name fields.

Swapping Specifications:

A slider labeled "Swapping Percentage(%)" is set to 10. The slider range is from 0 to 50, with major ticks every 10 units. A small box to the right of the slider displays the value "10".

Attributes	Swap	Fix	Differ	Other
Age	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
EmplType	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Educ	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
MarStatus	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Race	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Sex	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
AveHours	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Salary	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>

Sample Specifications File

```
num.records=1024
```

```
data.file=webswapdemo.orig
```

```
log.file=webswapdemo.log
```

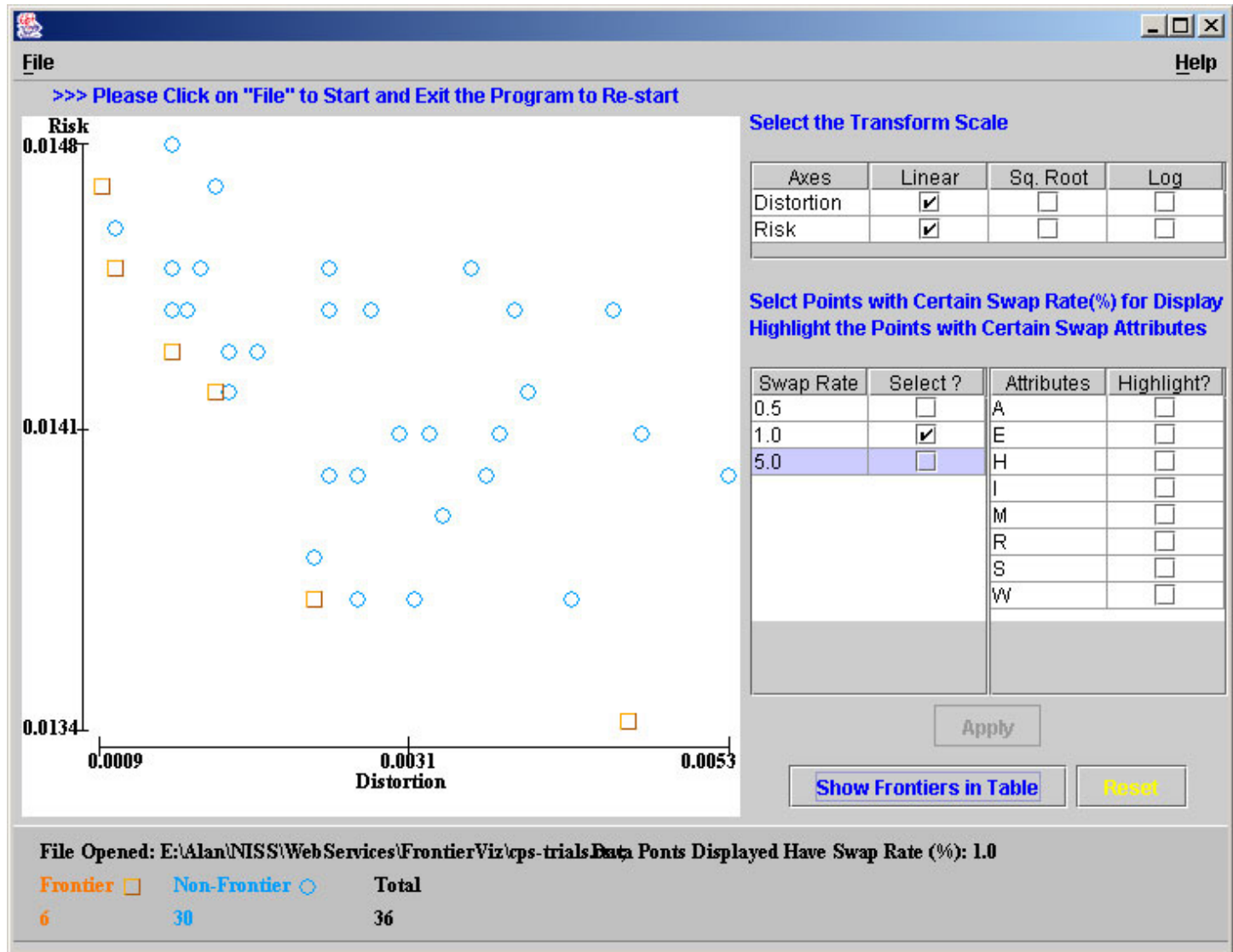
```
output.file=webswapdemo.swapped
```

```
swap.rate=0.25
```

```
attribute.specs=S,0,0,0,0,0,0,0
```

```
csv.type=MS
```

Frontier Visualization Tool



Sample Study: CPS-8

- Risk: Fraction of unswapped records in low-count cells in post-swap data

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

- Distortion: Hellinger distance between pre- and post-swap data tables
- 3 rates: 1%, 2%, 10%
- 36 choices of swap attributes
 - All single attributes
 - All pairs of attributes

Results of the Study

- 1% frontier = {AveHours, Educ, AveHours+Educ, AveHours+Sex, Sex, AveHours+MarStat, EmplType+MarStat}
- 2% frontier = {Educ, AveHours, Race, AveHours+Race, EmplType+Sex, EmplType+Race, AveHours+MarStat, Age+Income}
- 10% frontier = {Educ, Race, Educ+Race, AveHours+Race, EmplType+Race, Age+Race}

Future Research

- Databases with record weights
- Inference-based measures of distortion
 - Example: information loss in log-linear models
- Combining data swapping with other methods for SDL
 - Example: category aggregation
- DSTK: Data Swapping Toolkit

Downloads

- NISS WebSwap client, user documentation, sample data sets
 - www.niss.org/WebServices/dg/WebSwap.html
- Technical reports
 - www.niss.org/dg/technicalreports.html

References

- L. C. R. J. Willenborg and T. de Waal (2000). *Elements of Statistical Disclosure Limitation*. Springer–Verlag, New York.
- S. Gomatam and A. F. Karr (2003). Distortion measures for categorical data swapping. Submitted to *JOS*.
- S. Gomatam, A. F. Karr and A. P. Sanil (2003). A risk-utility framework for categorical data swapping. Submitted to *JOS*.
- A. P. Sanil, S. Gomatam, A. F. Karr and C. Liu (2003). NISS WebSwap: A Web Service for data swapping. *J. Statist. Software* 8(7).
- NISSWebSwap, Version 1.1, User Documentation. Downloadable at www.niss.org/WebServices/dg/NISSWebSwap_v11.pdf