**CASC PROJECT**

Computational Aspects of Statistical Confidentiality

# THE RANKSWAPPING ALGORITHM (SOFTWARE DOCUMENTATION AND RELATED PAPER)

**(WORKPACKAGE 1.1, DELIVERABLE 1.1-D5-bis)**

**Francesc Sebé Feixas, Josep Domingo-Ferrer, Josep M. Mateo-Sanz,**
**Antoni Martínez-Ballesté, Àngel Torres and Narcís Macià**
**Universitat Rovira i Virgili**
**{fsebe,jdomingo,jmateo,anmartin,atorres,nmacia}@etse.urv.es**

**11 June 2002**

This deliverable is intended to document the rankswapping software contributed in Deliverable 1.1-D6bis. The algorithm being implemented is described in the scientific paper attached to this deliverable.

The software in Deliverable 1.1-D6bis consists of a single piece of portable C++ code (rankswap.cpp). It can be compiled and run in any environment.

**<u>User's manual</u>**

Software usage:

```
$ rankswapping parameters_file
```

This program performs a random swapping between the elements of a column of data (see attached [Moore96]). Suppose that you have the following data:

| | | |
|---|---|---|
| A | A | A |
| B | B | B |
| C | C | C |
| D | D | D |
| E | E | E |

A possible output for the program could be:

| | | |
|---|---|---|
| E | D | B |
| C | A | D |
| D | E | E |
| A | B | C |
| B | C | A |

Thus, data are mixed up and records are “broken”. In spite of it, means and other statistics can be done on the rankswapped columns.

The data are kept in a text file, so called the **input file**. For example, the following is a toy input file:

```
10 4 6 2 1
12 3 7 1 2
17 2 5 1 3
21 2 8 2 4
9 3 3 3 5
12 4 7 3 6
12 4 6 3 7
```

In the above example, there are 7 records with 5 variables each.

The parameter file looks like this (note that comments can be included by adding lines starting with #):

```
####################################################
# Parameters example file for 'rankswapping'
# (Lines with '#' are comments, CR are also skipped)
#####################################################

# Input and output data files
input.dat
output.dat

# Number of records in data file
1080

# Number of columns (variables)
13

# Percent (p)
50
```

The parameters in this file must be described strictly in the following order:

- Input and output data files. If the output file does not exist, a new one will be created. Otherwise, the results are appended to the existing output file.
- Number of records. Number of rows of the input file data array.
- Number of columns. Number of columns (variables) are in the data array.
- Percent. It defines de maximum distance an element can be placed from its origin. For example, if it equals 50 and the number of records is 1000, the first element will be randomly placed into the 500th row at the most.

The output file will contain the rankswapped data.

**References (attached to this document)**

[Moore96] R. Moore (1996), ``Controlled data swapping techniques for masking public use microdata sets'', U. S. Bureau of the Census (manuscript).