Project N°: 262608

ACRONYM: Data without Boundaries

DELIVERABLE D11.5

(Software for New Masking Techniques
& for New Cell Suppression Method)

WORK PACKAGE 11

(Improved Methodologies for Managing Risks of Access to Detailed OS Data)

REPORTING PERIOD:	From: Month 36	To: Month 48
PROJECT START DATE:	1st May 2011	DURATION: 48 Months
DATE OF ISSUE OF DELIVERABLE:	Sept. 2014	
DOCUMENT PREPARED BY:	P3, P20	URV, ULL

Combination of CP & CSA project funded by the European Community
Under the programme "FP7 - SP4 Capacities"
Priority 1.1.3: European Social Science Data Archives and remote access to Official Statistics

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 262608 (DwB - Data without Boundaries).

PREFACE

This document is the descriptive part associated with the source code that is delivered as deliverable D11.5.

The first part of this descriptive report concerns the code provided by URV. This code is an R implementation of the New Masking Techniques.

The second part concerns the code provided by ULL. That code is a C implementation of a Cell suppression method, using Free and Open Source solvers.

TABLE OF CONTENTS

PART I - R IMPLEMENTATION OF NEW MASKING TECHNIQUES

INTRODUCTION

The software consists of a graphical user interface (GUI) programmed in JAVA, which implements all the methods included in the next chapters. In the Taxanonym package, the MicroHybrid [1] algorithm and the Data Shuffling [2] algorithm are implemented. Moreover, in the Kanonymdwb package, the MDAV_ID, MDAV_SWAP and IR_SWAP algorithms [3, 4, 5, 6] are implemented.

The GUI allows the user to choose the parameter *K* (in case of the microaggregation algorithms) and also the different kinds of variables (e.g. quasi-identifiers in MDAV_ID or MDAV_SWAP or confidentials in IR_SWAP).

The process is shown in the Logger on the right side of the interface.

Apart from this graphical program, two R-packages have also been implemented. In the following chapters their implementations and features are deeply explained. Although these packages are attached with the deliverable, both are in process to be downloadable through the CRAN official repository.

This document includes also the structure of the attached deliverable, and an easy how-to with instructions about the proper configuration of the packages into the R environment.

TAXANONYM PACKAGE

Introduction

The taxanonym package includes the functions *microhybrid* and *shuffledata*, which are the implementations of the MDAV + MicroHybrid perturbation and Data Shuffling algorithms. In this package, it is necessary to have installed a Java Virtual Machine (such as the classic JDK libraries) and the packages MASS and SDCMicro.

MicroHybrid Algorithm – Features

Description

Implements the MicroHybrid Algorithm. It is a combination of MDAV microaggregation of the dataset and the application of a synthetic algorithm on each one of the created groups. It preservs as much as possible the mean and variance vector, and moreover the covariance matrix of the perturbed data.

Usage

```
microhybrid(file, atts, hierarchy, k)
```

Arguments

`file`	The dataset file.
`atts`	The attributes file. It must have the following structure: att_name; category (nominal, ordinal, numeric); confidentiality (confidential, non_confidential)
`hierarchy`	The hierarchy file for the nominal attributes. The structure must be the same as in the attached example: root: son1, son2, son3...; son1: son1.1, son1.2.....;
`k`	The number of records per microaggregated group.

Details

IMPORTANT: The dataset file must include the header with the attributes' name. The other entry files must accomplished the conditions exposed above.

Value

The returned dataset will have the same structured as the original one, but filled with String values, in order to be exported or treated freely.

Author(s)

Guillem Rufian-Torrell, Universitat Rovira i Virgili.

References

J. Domingo-Ferrer, K. Muralidhar and G. Rufian-Torrell, "Anonymization methods for taxonomic microdata", Lecture Notes in Computer Science, Vol. 7556 (Privacy in Statistical Databases-PSD 2012), pp. 90-102, Sep 2012, ISSN: 0302-9743.

Data Shuffling Algorithm – Features

Description

The Data Shuffling Algorithm. First of all, it converts the nominal attributes of the set to marginalty-based ones, in order to compute the numerical transformations of the Algorithm. Afterwards, the original Shuffle Algorithm is applied.

Usage

```
shuffledata(file, atts, hierarchy, responses, predictors)
```

Arguments

`file`	The dataset file.
`atts`	The attributes file. It must have the following structure: att_name; category (nominal, ordinal, numeric); confidentiality(confidential, non_confidential)
`hierarchy`	The hierarchy file for the nominal attributes. The structure must be the same as in the attached example: root: son1, son2, son3...; son1: son1.1, son1.2.....;
`responses`	Attributes which will be the responses used in the Shuffling procedure. They must be separated with a ' + ' separator.
`predictors`	Attributes which will be the predictors used in the Shuffling procedure. They must be separated with a ' + ' separator.

Details

For further details, please visit the ?shuffle page or the published references.

Value

A dataset with only the response values shuffled will be returned.

Author(s)

Guillem Rufian-Torrell, Universitat Rovira i Virgili

References

Muralidhar, K. and R. Sarathy, "Data Shuffling- A New Masking Approach for Numerical Data," Management Science, 52(5), 658-670, 2006.

KANONYMDWB PACKAGE

Introduction

The kanonymdwb package includes the functions MDAV_ID, MDAV_SWAP and IR_SWAP, which are the implementations of different kinds of microaggregations, depending on the selected variables. In this package, it is only necessary to have installed a Java Virtual Machine (such as the classic JDK libraries).

MDAV_ID – Features

Description

Implement the MDAV_ID Algorithm. A microaggregation algorithm (MDAV) is computed only for the quasi-identifier attributes of the dataset. Afterwords, the respective values of the attributes are replaced by the centroid of each of the groups.

Usage

```
MDAV_ID(file, atts, hierarchy, QI, k)
```

Arguments

`file`	The dataset file.
`atts`	The attributes file. It must have the following structure: att_name; category (nominal, ordinal, numeric); confidentiality (confidential, non_confidential)
`hierarchy`	The hierarchy file for the nominal attributes (only necessary if there are hierarchical variables in the dataset). The structure must be the same as in the attached example: root: son1, son2, son3...; son1: son1.1, son1.2.....;
`QI`	The name of the quasi-identifiers. They must be separated by a ' + ' character.
`k`	The number of records per microaggregated group.

Details

IMPORTANT: The dataset file must include the header with the attributes' name. The other entry files must accomplished the conditions exposed above.

Value

The returned dataset will have the same structured as the original one, but filled with String values, in order to be exported or treated freely.

Author(s)

Guillem Rufian-Torrell, Universitat Rovira i Virgili.

References

J. Domingo-Ferrer, K. Muralidhar and G. Rufian-Torrell, "Anonymization methods for taxonomic microdata", Lecture Notes in Computer Science, Vol. 7556 (Privacy in Statistical Databases-PSD 2012), pp. 90-102, Sep 2012, ISSN: 0302-9743.

J. Domingo-Ferrer and J. M. Mateo-Sanz. Practical data-oriented microaggrega- tion for statistical disclosure control. IEEE Transactions on Knowledge and Data Engineering, 14(1):189-201, 2002.

J. Domingo-Ferrer. A survey of inference control methods for privacy-preserving data mining. In Privacy-Preserving Data Mining, volume 34 of Advances in Database Systems, pages 53â€“80. Springer, 2008.

P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Tech. rep. SRI-CSL-98-04, SRI Computer Science Laboratory, Palo Alto, CA, 1998.

MDAV_SWAP – Features

Description

Implement the MDAV_SWAP Algorithm. A microaggregation algorithm (MDAV) is computed only for the quasi-identifier attributes of the dataset. Afterwords, the respective values of the attributes are swapped with a permutation algorithm inside the groups.

Usage

```
MDAV_SWAP(file, atts, hierarchy, QI, k)
```

Arguments

`file`	The dataset file.
`atts`	The attributes file. It must have the following structure: att_name; category

(nominal, ordinal, numeric); confidentiality (confidential, non_confidential)

`hierarchy` The hierarchy file for the nominal attributes (only necessary if there are hierarchical variables in the dataset). The structure must be the same as in the attached example: root: son1, son2, son3...; son1: son1.1, son1.2.....;

`QI` The name of the quasi-identifiers. They must be separated by a ' + ' character.

`k` The number of records per microaggregated group.

Details

IMPORTANT: The dataset file must include the header with the attributes' name. The other entry files must accomplished the conditions exposed above.

Value

The returned dataset will have the same structured as the original one, but filled with String values, in order to be exported or treated freely.

Author(s)

Guillem Rufian-Torrell, Universitat Rovira i Virgili.

References

J. Domingo-Ferrer, K. Muralidhar and G. Rufian-Torrell, "Anonymization methods for taxonomic microdata", Lecture Notes in Computer Science, Vol. 7556 (Privacy in Statistical Databases-PSD 2012), pp. 90-102, Sep 2012, ISSN: 0302-9743.

J. Domingo-Ferrer and J. M. Mateo-Sanz. Practical data-oriented microaggregation for statistical disclosure control. IEEE Transactions on Knowledge and Data Engineering, 14(1):189-201, 2002.

J. Domingo-Ferrer. A survey of inference control methods for privacy-preserving data mining. In Privacy-Preserving Data Mining, volume 34 of Advances in Database Systems, pages 53-80. Springer, 2008.

P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Tech. rep. SRI-CSL-98-04, SRI Computer Science Laboratory, Palo Alto, CA, 1998.

IR_SWAP – Features

Description

Implement the IR_SWAP Algorithm. A microaggregation algorithm (MDAV) is computed for each one of the confidential attributes of the dataset. Afterwords, the respective values of the attributes are swapped with a permutation algorithm inside the groups.

Usage

```
IR_SWAP(file, atts, hierarchy, confidentials, k)
```

Arguments

`file`	The dataset file.
`atts`	The attributes file. It must have the following structure: att_name; category (nominal, ordinal, numeric); confidentiality (confidential, non_confidential)
`hierarchy`	The hierarchy file for the nominal attributes (only necessary if there are hierarchical variables in the dataset). The structure must be the same as in the attached example: root: son1, son2, son3...; son1: son1.1, son1.2.....;
`QI`	The name of the quasi-identifiers. They must be separated by a ' + ' character.
`k`	The number of records per microaggregated group.

Details

IMPORTANT: The dataset file must include the header with the attributes' name. The other entry files must accomplished the conditions exposed above.

Value

The returned dataset will have the same structured as the original one, but filled with String values, in order to be exported or treated freely.

Author(s)

Guillem Rufian-Torrell, Universitat Rovira i Virgili.

References

J. Domingo-Ferrer, K. Muralidhar and G. Rufian-Torrell, "Anonymization methods for taxonomic microdata", Lecture Notes in Computer Science, Vol. 7556 (Privacy in Statistical Databases-PSD 2012), pp. 90-102, Sep 2012, ISSN: 0302-9743.

J. Domingo-Ferrer and J. M. Mateo-Sanz. Practical data-oriented microaggrega- tion for statistical disclosure control. IEEE Transactions on Knowledge and Data Engineering, 14(1):189-201, 2002.

J. Domingo-Ferrer. A survey of inference control methods for privacy-preserving data mining. In Privacy-Preserving Data Mining, volume 34 of Advances in Database Systems, pages 53â€“80. Springer, 2008.

P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Tech. rep. SRI-CSL-98-04, SRI Computer Science Laboratory, Palo Alto, CA, 1998.

CONCLUDING REMARKS

Directory tree

The deliverable offers the following structure, according to the structure that all the R-packages must obey.

Root

- ***Java Program of the deliverable.***
- ***R Packages(taxanonym & kanonymdwb)***
    - ***Src****: Contains the java binaries (black box) that run the microaggregation functions.*
    - ***Datasets****: Example datasets (with their respective hierarchy and attribute files) that can be used to run tests.*
    - ***Man****: Help files.*
    - ***R****: Source R files. Please to do not modify them, in order to avoid future execution problems.*
- ***Example Datasets:*** *They than can be run with both the GUI and the R-Packages*.

How-To-Install the R-Packages

1. Download and install R. Sources, binaries and documentation for R can be obtained via CRAN (http://www.r-project.org). It is also advisable to download R-Studio (http://www.r-studio.com).
2. After installation of the base package, run R.
3. Be sure that the JAVA libraries are installed in the computer. If not, download the most recent JDK-Library at http://www.oracle.com/technetwork/java/javase/downloads/index-jsp-138363.html
4. Install packages *MASS* and *sdcMicro*. They are essential to execute the *Taxanonym* functions. You can do it using GUI or using the next R command:
    - install.packages(c("MASS", "sdcMicro"))
5. Install the new *Taxanonym* and *kanonymdwb* packages. You can do it using the GUI option "*Install package*". Browse and select file *taxanonym.tar.gz* for *Taxanonym* and *kanonymdwb.tar.gz* for the *kanonymdwb* one.

6. Load packages *Taxanonym* and *kanonymdwb*. You can do it using GUI or using the next R command.
    - library("Taxanonym")
    - library("kanonymdwb")

Now you can use the *microHybrid, shuffleData, MDAV_ID, MDAV_SWAP* and *IR_SWAP* functions. You can see their respective documentations by executing the R command *help("function")*.

The zip file also contains dataset examples to learn how to use the functions and the correct structure of the input files. As the Working Directory is set by default to point the external sources, it's extremely recommendable that the datasets are placed in the "*data"* folder.

REFERENCES

[1] J. Domingo-Ferrer, K. Muralidhar and G. Rufian-Torrell, "Anonymization methods for taxonomic microdata", Lecture Notes in Computer Science, Vol. 7556 (Privacy in Statistical Databases-PSD 2012), pp. 90-102, Sep 2012, ISSN: 0302-9743.

[2] Muralidhar, K. and R. Sarathy, "Data Shuffling- A New Masking Approach for Numerical Data," Management Science, 52(5), 658-670, 2006.

[3] J. Domingo-Ferrer and V. Torra. A quantitative comparison of disclosure control methods for microdata. In P. Doyle, J.I. Lane, J.J.M. Theeuwes, and L. Zayatz, editors, Confidentiality, Disclosure and Data Access: Theory and Practical Ap- plications for Statistical Agencies, pages 111-134. North-Holland, Amsterdam, 2001.

[4] J. Domingo-Ferrer and J. M. Mateo-Sanz. Practical data-oriented microaggrega- tion for statistical disclosure control. IEEE Transactions on Knowledge and Data Engineering, 14(1):189-201, 2002.

[5] J. Domingo-Ferrer. A survey of inference control methods for privacy-preserving data mining. In Privacy-Preserving Data Mining, volume 34 of Advances in Database Systems, pages 53-80. Springer, 2008.

[6] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Tech. rep. SRI-CSL-98-04, SRI Computer Science Laboratory, Palo Alto, CA, 1998.

PART II - C IMPLEMENTATION OF CELL SUPPRESSION METHOD

INTRODUCTION: PURPOSE OF SOFTWARE

During the first two years of the DwB project we have being working on the implementation of computer programs as described in the WP11 tasks assigned to our team at ULL.

We have produced the following codes:

[CSP] A computer program to use Cell Suppression on tables.

Cell Suppression is a classical methodology to protect private information when publishing statistical tables. It is based on suppressing some cell values. The cells containing private information must be suppressed, but potentially also others may also be suppressed too. The combinatorial problem of selecting the other cells is known to be extremely complicated, and for that reason we have sophisticated programs are necessary to efficiently apply this methodology in practice.

We have implemented a new program based on a previous code already integrated in tau-ARGUS. The main contribution is that the new program can be linked to a non-commercial library to solve mathematical programming models. This is the main aim of our task in WP11. The new program keeps the same algorithm of the previous implementation, and also it keeps the possibility of using the commercial libraries (Cplex and Xpress) as before. But now, in addition, it can use alternatively SCIP, which is under the ZIB academic license.

CONTEXT

We have presented our research results in different forums, including

- A seminar at "Statistics and Operational Research Doctoral Training Centre", Lancaster University, 29 June 2012.
  Title: "Statistical Disclosure Control: Techniques to protect confidential information "
  Web: http://www.stor-i.lancs.ac.uk/event-info/stori-seminar-gonzalez

- A seminar at Westat Inc (Washington), 26 November 2012. Title: "Statistical Confidentiality: Modern Techniques to Protect Sensitive Cells when Publishing Tables"
  Web: http://washstat.org/sem2012.html

HOW IT WORKS

A compressed file containing all source codes and documentation is available in https://www.webs.ull.es/users/jjsalaza/public/ECTA&CSP.zip

The compressed file contains several folders, for the CSP-code as described above as well as for the CTA-code developed previously as deliverable D11.4.

Each folder contains a README file describing all the details of the code, including how to compile it, what it does, examples, etc. It is fully documented. It also contains two subfolders: one with the source code and another with the executable. The executable contains an instance to test that it runs properly. The source code allows a computer programmer to change the program and produce a new executable or a DLL library to be called from another software. All files have been created using non-commercial tools, so no commercial software is necessary to read, edit or compile them.

AVAILABILITY TO WIDER AUDIENCE

All the material is publicly available at
https://www.webs.ull.es/users/jjsalaza/public/ECTA&CSP.zip

The author of the material is Juan José Salazar Gonzalez.
The owner is University of La Laguna.

It can be used in public and private organizations, even distributed, but it cannot be commercialized by any organization without previously a written agreement with the owner.

It is subject to the ZIB Academic License (http://scip.zib.de/academic.txt), like SCIP.

The source code and documentation may be made available for use by third parties upon request to and formal approval of the owner only.