My Three-day Encounter with ARGUS: A Report for IASSIST

by Chiu-Chuang (Lu) Chou [1]

Have you heard of the story about Argus in Greek mythology? He was a giant with one hundred eyes. Hera, Zeus's wife, assigned him to guard Zeus's mistress Io. Eager to get Io back, Zeus sent Hermes to kill the giant. Argus was then transformed into a peacock.

ARGUS is also the name for a practical tool to guard data. It has been developed by the Computational Aspects of Statistical Confidentiality (CASC) project, which is part of the Fifth Framework of the European Union. ARGUS was developed under Windows NT and runs under Windows versions from Windows 95. It intends to "modify unsafe data in such a way that safe (enough) data emerge, with minimum information loss,"[2] so that data producers can safely release the data to researchers and the public. ARGUS has two components in achieving the statistical disclosure control (SDC). μ-ARGUS is for safeguarding microdata, while τ-ARGUS is designed to make tabular data safer.

As a data librarian at a large research-oriented university, I am well aware of the confidentiality issues related to social science data sets. Since the Data and Program Library Service, where I work, only houses and disseminates public use data sets, I am interested in what the principal investigators need to do before they can safely release their data. In this report, I will summarize what I have learned in a three-day workshop called Statistical Disclosure Control for Data Confidentiality.

This workshop was hosted by the Center for Demography of Health and Aging (CDHA) from November 10-12, 2004 at the Pyle Center in the University of Wisconsin Madison. Anco Johannes Hundepool, Eric Schulte Nordholt, and Peter Paul de Wolf, three specialists from Statistics Netherlands, were the instructors. Their lectures and exercises covered the mathematical aspects of statistical disclosure control and the application of these methods in the ARGUS software. Participants came from the National Bureau of Economic Research, the Bureau of Labor Statistics, National Center for Health Statistics, and research centers at the University of Wisconsin, the University of Michigan and the University of Pennsylvania.

Why is Statistical Disclosure Control important?

Statistical Disclosure Control has attracted much attention recently because complex statistical analysis can easily be done on powerful PCs. In addition, the ease of linking files from different data sources presents a real threat of re-identifying business entities or individuals from public use microdata and tabular data. To balance the stricter legal regulations and increased data needs from policy makers, several statistical agencies[3] in the European Union have taken on the challenges in designing better SDC methods and building a new tool to balance the need for data and the need for confidentiality protection. The Computational Aspects of Statistical Confidentiality (CASC, http://neon.vb.cbs.nl/casc/) project is the result of this collaboration from 2001 to 2003. The CASC project comprises not only statistical theories and methods, but also ARGUS software development.

Statistical agencies and other data collectors have always removed direct identifying variables, like names, addresses, and social security numbers, before they released their data. However, such conventional practices during data processing are no longer adequate to protect respondents in the current computing world. Rare combinations of indirect/non-sensitive identifiers can re-identify certain respondents in microdata. Reducing these disclosure risks is very important before the release of microdata.

Data producers can apply various statistical methods and risk models to make their data safe using statistical packages like SAS, SPSS and Stata. However, it is very time consuming to do the global recoding and case swapping with any existing statistical packages. Meanwhile, the data producers need to document any changes they have made to the original data to meet their SDC criteria. As one can imagine, it is a substantial task to produce a safe data file. To address the needs of SDC in producing public-use data and to build an efficient tool to apply SDC was the main goal of the CASC project. ARGUS is the SDC application derived from the collaboration of many CASC researchers. It was first written in Borland C++ and then converted to Visual C++ with its user interface written in Visual Basic.[4]

μ-ARGUS: a SDC tool for creating safer microdata files
The current version of *μ*-ARGUS can read ASCII data files in fixed format, free format with a defined separator or free format with variable names in the first line. Users can provide a metadata description file or use *μ*-ARGUS to specify the metadata interactively. Value lists of the variables can be supplied as external files or entered as metadata attributes. *μ*-ARGUS will identify the records at risk by checking frequency tables of combinations of identifying variables. Low frequencies are considered a risk of re-identification. After the metadata and data file are read in to *μ*-ARGUS, users can specify the set of tables manually or use one of the two basic rules used in Statistics Netherlands for producing microdata files for researchers and for public use files. When users are satisfied with the tables, they press the button "Calculate tables" and *μ*-ARGUS will calculate the frequency tables automatically.

After the tables are calculated, users can start Disclosure Control in *μ*-ARGUS. You will select those variables that are identified as posing dangers of re-identification and apply various methods, such as recoding variables, suppressing values, perturbing values, applying top and bottom coding, adding noise, masking, PRAM (post-randomisation) and micro-aggregation to bring them to a safe level. If the result is a file with too much information loss, users can easily go back to the original file and apply a different risk model to reduce the level of information loss. Any changes made to the original data are documented for future reference, so users can examine the log files and see what risk models have been applied and how the data have been changed. When finally a safe file is generated, it can be output as an ASCII file with an accompanying metadata file.

τ-ARGUS: an SDC tool to publish safe tabular data
It is a misconception that aggregated data such as tabular data is safe. There are risks of disclosure if aggregated tables are not constructed with statistical disclosure control methods. In general the cells in a table should not be "too small" to disclose confidential information. All statistical agencies probably have their own rules for the minimum safe values for cells in their released tabular data. In addition, they need to detect when the information in aggregated tables is not just statistics but has the risk of group disclosure.

τ-ARGUS, like its twin *μ*-ARGUS, has SDC methods built in to facilitate evaluation of tabular data to see if they are safe. τ-ARGUS includes many sensitivity measures, such as a minimum number rule (threshold rule), an (n, k) dominance rule, a p% rule and a p/q rule (prior-posterior rule) to check the disclosure risks for magnitude tables. When a large number of sensitive cells are present in a table, data producers can use the Table Redesign feature in τ-ARGUS to combine rows and columns to eliminate the sensitive cells. Other methods such as suppression or rounding techniques can also make those cells safe. When a series of tables are created from the same microdata source, τ-ARGUS can effectively perform SDC on all these tables, so they can all be protected in one single session instead of several sessions of SDC. τ-ARGUS has four output options for writing out safe tables. They are CSV-format, CSV for pivot table, text file with code-value, or intermediate format.

ARGUS software and users' manuals are freely available from the CASC web site, http://neon.vb.cbs.nl/casc/. Users can click on *μ*-ARGUS and τ-ARGUS links on the side menu to download the most current version of this tool.

Legal Issues and Practices Pertaining to the Netherlands
Eric Schulte Nordholt gave a report on the legal issues related to SDC in the Netherlands and the European Union. He covered the ethical codes in several statistical organizations, laws and statutes in the Netherlands and how they affect the way Statistics Netherlands distributes their microdata and tabular data.

Even with well-implemented SDC, certain sensitive data sources are still unsafe to be released as public-use microdata or tabular data. However, researchers and policy makers need access to the data to conduct their studies. To balance the protection of confidentiality and the need for sensitive data, Statistics Netherlands has set up the Center for the Research of Economic Microdata (CEREM, http://www.cbs.nl/en/service/research/cerem/). This on-site data center provides researchers access to enterprise data in the Netherlands. Researchers need to comply with a set of strict rules before they can use the restricted data in the center. The on-site room is equipped with a stand-alone PC without e-mail, Internet, or any external drives. All the prospective publications will be screened for safety.

In 2002 the Center for Policy Studies was established to provide ministries with optimal statistical information. In addition to an on-site data center, researchers can submit their scripts remotely to be executed. First the job is run on test datasets and errors are corrected. The final script is then executed on real data and the results are sent to the researchers. At the start of their projects, researchers have to take an online course on SDC offered by the Center for Policy Studies. Any subsets created in the on-site data center or obtained via remote execution have to go through a safety check. It is labor intensive but necessary. To make this job easier, a new tool, ϱ-ARGUS has been developed and is being tested now.

Current User Base of ARGUS Software
Since ARGUS is a fairly new tool for Statistical Disclosure Control (SDC), it is mainly used in the national/state statistical offices among those countries involved in the European Union's the Computational Aspects of Statistical

Confidentiality (CASC) project. However, Anco Johannes Hundepool, Eric Schulte Nordholt, and Peter Paul de Wolf, the three specialists at Statistics Netherlands have conducted several ARGUS workshops to promote this tool and its SDC methodology. Their latest workshop was given on April 13 and 14, 2005 in Sydney Australia at the 55th Session of the International Statistical Institute. It is likely that they will give their workshop in future IASSIST annual conference.

Statistical Disclosure Limitation practices in the U.S. statistical agencies

Unlike the Netherlands, the U.S. Federal statistical system is not centralized and is comprised of over 70 agencies according to an Office of Management and Budget (OMB) report[5]. So how do these agencies protect the confidentiality of data that they collect? What SDC methods are used by Federal statistical agencies before they disseminate their public use microdata and tabular data?

The interagency Federal Committee on Statistical Methodology (FCSM) was established in 1975 to recommend standards for statistical methodology to be followed by federal statistical agencies. FCSM investigates problems which affect the quality of Federal Statistical data, as well as makes suggestions for improving statistical methodology in federal agencies. The FCSM has about twenty members. This network of Federal agency personnel has focused primarily on data quality. It has published a Confidentiality and Data Access Committee (CDAC) Checklist (http://www.fcsm.gov/committees/cdac/checklist_799.doc). This list consists of a series of questions that can assist an agency's Disclosure Review Board to determine the suitability of releasing either public use microdata files or tables. Please note that FCSM uses the term Statistical Disclosure Limitation (SDL) or Statistical Disclosure Restriction (SDR), not Statistical Disclosure Control, when it discusses different statistical disclosure techniques. Another important document is FCSM's Statistical Policy Working Paper # 22 (SPWP # 22): Report on Statistical Disclosure Limitation Methodology (http://www.fcsm.gov/working-papers/spwp22.html). It provides 12 recommendations to improve disclosure limitation practices. These two documents are the viable foundation for SDL practices in federal statistical agencies.

Similar to Statistics Netherlands, U.S. Federal statistical agencies have developed their own procedures to provide researchers access to their sensitive data. These procedures can be classified into three categories: on-site research centers, remote access, and data use agreements or licenses.[6] Three examples follow.

In 2003, the Census Bureau's Center for Economic Studies developed and opened several Research Data Centers (RDCs) around the country. The RDCs provide a secure Census Bureau environment where researchers may have limited access to confidential economic and demographic microdata, with appropriate safeguards to protect data confidentiality. Researchers need to submit their proposals to the RDCs first. After their research projects are approved, they will pay for the costs associated with the work, such as computer charges. At each RDC site, stand-alone workstations without removable media and network connections are set up in a secured and locked room. All the researchers' materials will be inspected before they are removed from the RDC. Disclosure reviews are performed on the researchers' output.

The National Center for Health Statistics (NCHS) has a remote access system for researchers in addition to an RDC in their headquarters in Hyattsville, Maryland. After their proposals are approved, researchers can submit their work electronically to staff at the NCHS' RDC. All submitted programs are reviewed for non-allowed commands, such as PROC TABULATE or PROC IML in SAS. All output goes through SDL review before they are sent back to researchers.

In the third case, the National Center for Education Statistics (NCES) licenses their restricted data to researchers for them to use at their home institutions. In their formal letter of data request, researchers need to specify their research scope and the time period for the loan of the restricted files. They also need to provide a security plan compiled with NCES's requirements. Each data user of the restricted files is required to sign an affidavit of nondisclosure. NCES conducts unannounced, unscheduled inspections of the licensee's site to assess compliance with the provisions of the license, security procedures, and the licensee's submitted security plan. Any violation subjects the licensee to immediate revocation of the license by NCES, or a report of the violation to the U.S. Attorney. The restricted data files need to be returned to NCES upon the completion of the project.

The Confidentiality and Data Access Committee (CDAC) web site (http://www.fcsm.gov/committees/cdac/cdac.html) has a link to Resources for Confidentiality and Data Access. It lists many important papers and reports for people who are interested in the topic.

Epilogue

The workshop announcement had proclaimed: "Workshop materials and presentations will be most accessible to those with graduate training in statistical methods (e.g., econometrics, demographic methods) and researchers experienced in the quantitative analysis of panel or longitudinal survey data." I was a bit concerned about how accessible those materials would be to a data librarian, like me, who has no formal training in either statistics or research methods. So with a curious mind, I went to the workshop and sat through all three days of lectures and

exercises even though I did not understand any of the intimidating mathematical formulas.

I was very impressed by how intuitive the ARGUS interface is. Users with the appropriate statistics background can easily make informed choices among built-in SDC methods and create a safe file for distribution. Lacking any statistical training, I am not in a position to appraise the built-in risk models and SDC methods in ARGUS. Yet, this workshop convinced me of the importance of SDC. My plan is to spread the SDC messages and share ARGUS on my campus. I hope that you will find my report on this workshop useful. To learn more about SDC and ARGUS, please visit the CASC web site, http://neon.vb.cbs.nl/casc/. It has links to many research papers relevant to the development of ARGUS software and the SDC theories and methods that are applied in ARGUS.

Endnotes

[1] Chiu-Chuang (Lu) Chou, Senior Special Librarian, Data and Program Library Service, University of Wisconsin Madison, United States. Email: cchou2@wisc.edu.

[2] Leon Willenborg and Ton de Waal, *Elements of Statistical Disclosure Control,* Lecture Notes in Statistics 105 (New York: Springer-Verlag, 2001).

[3] CASC project team includes Statistics Netherlands, Istituto Nationale di Statistica (Italy), University of Plymouth (UK), Office for National Statistics (UK), University of Southampton (UK), The Victoria University of Manchester (UK), Statistisches Bundesamt (Germany), University La Laguna (Spain), Institut d'Estadistica de Catalunya (Spain), Institut National de Estadisica, TU Ilmenau (Germany), Institut d'Investigacio Intelligencia Artificial-CSIC (Spain), Universitat Rovira i Virgili (Spain) and Universitat Politecnica de Catalunya (Spain).

[4] Anco Hundepool, "The ARGUS-software"(paper presentation, UN-ECE/Eurostat worksession, Luxembourg, April 7-9, 2003)..

[5] U.S. Office of Management and Budget, "Statistical Programs of the United States Government: Fiscal Year 2004," http://www.whitehouse.gov/omb/inforeg/04statprog.pdf (accessed on February 10, 2005)

[6] Virginia A. de Wolf, "Issues in Accessing and Sharing Confidential Survey and Social Science Data," Data Science Journal 2, (2003).