

Working Paper No. 30
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN COMMUNITIES (EUROSTAT)**

Joint ECE/Eurostat work session on statistical data confidentiality
(Luxembourg, 7-9 April 2003)

Topic (ii): New data release techniques

**IMPLEMENTING STATISTICAL DISCLOSURE CONTROL FOR
AGGREGATED DATA RELEASED VIA REMOTE ACCESS**

Contributed paper

Submitted by the National Statistical Institute of Italy¹

¹ Prepared by Luisa Franconi (franconi@istat.it) and Giovanni Merola (gmmerola@istat.it).

Implementing Statistical Disclosure Control For Aggregated Data Released Via Remote Access

Luisa FRANCONI and Giovanni MEROLA

ISTAT, Servizio MPS

Via C. Balbo, 16. 00184, Roma, Italy

e-mail:{ franconi, merola}@istat.it

Abstract: In this paper we give an overview of various approaches to the implementation of statistical disclosure control to tabular data released through the Web. We consider three generic groups of statistical disclosure control methods: source data perturbation, output perturbation and query-set restriction. Considering different types of Web-sites and implementation approaches we discuss the appropriateness and effectiveness of such statistical disclosure control methods.

Keywords: Remote access, tabular data, statistical disclosure control, on-line database

1. Introduction

Dissemination is one of the missions of National Statistical Institutes (NSIs); it is a way for giving society useful information and it also a way of motivating respondents to answering surveys. If NSIs have the legal standing to collect and release information, they also have the obligation to protect confidentiality of respondents, which sets a constraint on the amount of information that can be released. This trade-off is addressed by statistical disclosure control (SDC), which consists of a collection of techniques that make difficult matching confidential information with the identity of respondents from a set of data. SDC methods employ a variety of techniques that either alter or suppress some of the data released.

We classify the data released in three categories:

- **Microdata:** files with individual observations;
- **Tables:** total values carried by individuals that fall in given classifications;
- **Other statistics:** summaries of different types, for example: regression coefficients, relative indices, correlation coefficients, etc.

In this paper we do not consider the release of microdata files but only of aggregated values. In particular, we focus on the release of tables, although some of the conclusions that we draw can be applied to the release of other statistics, as well.

NSIs have a long tradition in publishing periodically printed reports of the data that they collect, which are sold or distributed freely. However, the Internet is becoming a standard channel with which public institutions communicate with general public. Institutions expect people to look for information on the Web and Internet navigators expect to find it on the Web. Furthermore, Web-sites are an ideal tool for disseminating data, as they are cheap, flexible, easy to update and accessible by most users (Blakemore, 2001). In fact, most NSIs provide on-line data in tabular form. Such Web-sites, to which we refer as Web-based Systems for Data Dissemination (WSDDs), require automated systems that release data upon request. WSDDs can be designed in many different ways, whether giving access to a predefined set of tables or allowing users to query any table choosing from a set of available variables. Therefore, SDC methods must be applied to WSDDs according to their structure and flexibility as well as to the type, quality and level of detail of the information released. In many cases SDC must also be combined with electronic access control.

We argue that SDC methods can be applied to WSDDs in two ways: *a priori*, that is, before releasing the tables, or *a posteriori*, that is, after the user has made his/her particular query. We refer to the former as PRE SDC and to the latter as POST SDC. According to this classification we comment on benefits and drawbacks of different SDC approaches. In Section 2 we present the essential framework of SDC

methods for the release of tabular data. In Section 3 we classify different SDC methods according to the way they are implemented in WSDDs and analyse the consequences of their application from the point of view of both the producer and the user. Finally, in Section 5 we present our conclusions casting the various WSDDs in the perspective of the general framework of lack of access versus information loss.

2. Statistical Disclosure Control for Tables

Tables have always been the essential data products of any NSI or statistical agency. Tables are built from the source data file containing records on individuals, called microfile, aggregating the values of the responses for the units falling in the categories of the chosen classifying variables. When the response is equal to one for all observations, the resulting table is a frequency table that gives the number of units in each cross-classification. As customary in SDC theory, we restrict the attention to tables for nonnegative responses thus including frequency tables.

SDC theory deals with releasing data without confidential information being traced back to respondents. To provide context to our discussion we briefly describe the framework of SDC.

A cell of a table is considered at risk of disclosure, or *sensitive*, if pre-defined rules are not satisfied (Willenborg and de Waal, 2001). The most used rules are: minimum number of units in each cell, called *threshold rule*, and maximum percentage of concentration (applicable to continuous variables), called *dominance rule*. These rules are based on the idea of impeding to an intruder estimating too closely confidential values.

Tables containing cells at risk can be protected by different techniques, which either inhibit the access to part of the information or distort the information released. Clearly, there exists a trade-off between the level of protection achieved for the data and the quality of the information released. Duncan *et al.*, 2001, propose a model to evaluate, also in a graphical way, this trade-off.

SDC methods for tabular data are either based on data transformation — input data masking, cell perturbation- or data suppression (see Willenborg and De Waal, 2001). The former may suffer from bias in the information released whereas the latter may prevent the release of information. The considerable effort gone into developing better and more targeted SDC methods, that effectively protect the confidentiality of respondents, increased the quantity of high standard products that can be safely released, with benefit for legitimate users. Effort has also gone into developing software for the application of methods that require intense computation; for example, the CASC project², funded by the European Union, developed two programs, ?-Argus and ?-Argus (Hundepool, 2001) that use sophisticated routines for the application of SDC for tabular data and individual data, respectively.

Duncan (2000) classifies SDC methods in terms of *disclosure limiting masks*, we regroup this classification according to the following three, broader, basic approaches, which can be used combined together or alone:

Perturbing the source data: records in the source data are *perturbed* before any data product is released. Such perturbations can be of different, nonexclusive, kinds: suppressing records, swapping some values between similar records (Dalenius and Reiss, 1982), applying Markov perturbation, (Gouweleeuw *et al.*, 1998) or model based perturbation (Franconi and Stander, 2002), adding random noise (Brand, 2002) or (sub-)sampling from the entire source data file, assigning wider categories to classifying variables. All these techniques allow the protector for a large degree of decision but it is often difficult to evaluate the level of protection achieved. Furthermore, several studies have shown (*e.g.*, Winkler, 1998 and Brand, 2002) that data protected by these techniques are often severely distorted.

Perturbing cell values: some or all values to be released are perturbed, either by adding random noise or by rounding them. Perturbative methods present the same drawbacks as those perturbing the source data.

² <http://neon.vb.cbs.nl/rsm/casc/>

Moreover, the resulting tables may be nonadditive (i.e. with marginal values not congruent with the inner cells values).

Suppressing cell values: sensitive cell values are not released. Together with these values also other nonsensitive cells must be suppressed, in order to avoid recovering sensitive values by differencing, this is the so called *complementary suppression*. Cell suppression is the most popular protection method in SDC, also because it can be easily customized with respect to a given loss function, differently from other methods. The algorithms for choosing the cells to be suppressed in an optimal way are complex and slow. There is an extensive literature in this area treating both heuristic and exact solutions (see, for example, Willenborg and de Waal (2001) and references therein). The drawback of cell suppression is the reduction in information released. Conceptually, the extreme case of cell suppression is the suppression of the whole table; while this practice leads to a great loss of information released, it can be convenient when the number of tables to be protected is large because it avoids the demanding computations for finding the complementary suppressions.

The above approaches can be applied also to protect the releases of other statistics, however, their application to specific quantities is still under study. For a more comprehensive review of SDC methods for tabular data see, for example, Duncan *et al.*, (2001), while for a detailed account see Willenborg and De Waal, (2001) and references therein.

Different methods can be adopted for the protection of data to be released and there will not be agreement on which methods are better. In the next section we consider the application of the different SDC methods to WSDDs, without discussing the merits of the methods themselves, but having in mind the peculiar problems posed by their application to automated release systems.

3. Problems peculiar to the application of Statistical Disclosure Control to WSDDs

WSDDs are Web-sites, accessible through the Internet, in which navigators can query tables automatically built from nonaccessible source data. Usually, WSDDs are set up for records measuring several classifying variables and response variables. Therefore, for each response there exists a high dimensional table formed using all the classifying variables, of which lower dimensional marginal tables are released. The most informative WSDDs allow users to query tables at their choice, choosing among combinations of classifying and responses variables contained in the source data. We call these sites *dynamic* as opposed to *static* ones, which link users to only a pre-established subset of all possible marginal tables. More sophisticated WSDDs that also offer other statistics, such as, for example, correlation or regression coefficients, will be referred to as *Virtual Laboratories* (V-Labs).

From the point of view of SDC, static WSDDs are not really different from printed releases, therefore standard SDC techniques can be applied for their protection. Methods that require intensive computations (i.e. partial suppression), however, may not be applicable when the number of tables to be protected is large. SDC for dynamic sites is more difficult for two main reasons: the information retrievable consist of a large number of linked tables³; the total information released to each user is different, cumulative and not known in advance. V-Labs are a hard challenge for SDC and both their design and protection are still under study. In the following we will focus mainly on the application of SDC to WSDDs that release tables but some of the results are applicable to V-Labs, too.

The application of SDC techniques to WSDDs is a difficult task, mainly for four reasons: methods and criteria must be standardized and implemented in an automatic system; usually there is a large number tables of high dimensions to be protected; some of the tables are *linked* that is, share some spanning variables and, when users can choose freely which tables to see, the total information queried by a user is not known beforehand. As regards the choice of the SDC method for protecting the data, it is evident that the more information is granted to users, the harder it is to control disclosure of confidential information.

³ Assuming that p classifying variables and q responses are offered for tabulation, the number of possible tables is $q2^p$.

Thus, SDC techniques must be tailored for each site with respect to the nature of the data and the information offered. Often, summaries of the data released in a WSDD have already been published (for example in printings), in this case WSDDs provide additional information and SDC techniques must be consistent with what has already been released. Furthermore, some WSDDs release data from databases that are updated with new data over time; in this case, extra care must be put in SDC for these WSDDs, because it must also be consistent through time.

As mentioned above, protecting tables released by WSDDs by suppressing few inner cells is not usually feasible, because of the demanding computations needed for each table. Rather, more realistically, the suppression of complete tables can be applied. We will refer to this approach as restricting the allowed query-set. Query-set restrictions are rules that evaluate whether a table can be released or not because it would be too risky. We reserve a special category to this approach because it is typical of WSDDs. Restrictions require auditing, (Malvestuto and Moscarini, 1999) and can be of different types: on the number of tables already released, on the dimension of the tables queried, on the goodness of fit of the tables unreleased that they allow, etc. (*e.g.* see Adam and Wortmann, 1989 and Fienberg, 2000, for a general review of different approaches). We consider the following generic approaches to protecting releases of a WSDD:

- (1) perturbing the source data (sampling, swapping, adding noise, etc.);
- (2) perturbing the output (adding noise, suppressing values, rounding, etc.);
- (3) restricting the allowed query-set (denying tables).

One way of limiting the possibility of breaching confidentiality through WSDDs is by restricting the access to registered users and releasing the output directly to them. In this way, WSDD administrators reduce the probability of vicious intruders accessing the data, regardless the type of data. In sites with restricted access, users must register and login is allowed only to those fulfilling given requirements. Requirements can have different nature, such as: the most generic one being “having a valid email address”; another common, stricter, one is “belonging to a research institution”. Registration may or may not include the electronic signature of a confidentiality agreement. In any case, tight access restrictions create a severe reduction in the publicness of the site, while looser restrictions are often not effective because of possible phoney email addresses. Another limitation to breaches of confidentiality can be obtained by releasing the output by email. In this case, users must provide a valid email address and sites administrators have records of what has been released and to whom. Another possibility for having records is to keep logs of IP’s connecting to the site, although this method is usually hidden and therefore harder to use as evidence. These restrictions, however, pertain to the security of the site and not to statistical disclosure control, hence will not be analysed here, although they are used in combination with SDC. An interesting analyses of access restrictions from SDC point of view can be found in David, 1998, and Blakemore, 2001.

4. Strategies for the application of SDC to WSDDs

SDC for WSDDs can be applied following two strategies: before queries are submitted (that is, before putting the data on-line) or after. We will refer to the former as PRE SDC and to the latter as POST SDC. PRE SDC is applied off-line, POST SDC can be carried out on-line, *on-the-fly*, but it can also be applied off-line, delaying the release of the output. The advantage of POST SDC is that it can be designed to be adaptive to the information previously released. If it is reasonable to assume that users do not co-operate among themselves, POST-SDC can be adapted to the information released to each user; in this case, non-anonymous access, auditing of users’ activity and, sometimes, logging of all releases, are required. Several sites that apply PRE SDC can be found on the Web such as, for instance: Italian Foreign Trade statistics at ISTAT⁴. POST SDC is still not as common as PRE SDC, a site that applies POST aggregation of geographical units is the NASS⁵ site that releases statistics on the usage of pesticides in

⁴ COEWEB <http://www.coeweb.istat.it/>

⁵ developed at NISS <http://niss.cndir.org/>, more information at <http://www.niss.org/dg/nass-system.html>

the US. Some sites use both PRE and POST SDC, such as, for example, the American Fact Finder⁶ (AFF). Next, we will briefly discuss the application PRE and POST of SDC to WSDDs.

WSDDs can be designed in many different ways, allowing for different amounts of information to be released. Dynamic WSDDs require software that builds the tables on-the-fly, possibly with embedded SDC. In any case, SDC techniques must be tailored to each WSDD in order to release the maximum possible information preserving privacy. In fact, researchers are developing fully automated expert systems for applying SDC to WSDDs (see, for example, Keller-McNulty and Unger, 1998, Malvestuto and Moscarini, 1999 and Fienberg, 2000, for theoretical treatment and Karr *et al.*, 2001, Zayatz, 2002 and Karr *et al.*, 2002, for examples of applications). Next, we discuss the PRE and POST application of the classes of SDC methods given above.

Perturbation of the source data

The perturbation of the source data is usually applied PRE, so that data are perturbed off-line and then the tables queried are built on these perturbed data. This approach is often applied to large data-sets, in which the number of units is large enough to allow for a consistent reduction or for the law of large numbers to apply. A common practice is that of building the tables from a small sub-sample of the complete data-set, as done, for example, at the AFF and for the 1996 Brazilian Census Data⁷. Also POST perturbation of the source data is possible, for example by extracting a new sub-sample for each query. Such practice seems reasonable because it can be implemented to be adaptive to queries, that is, for example, adapting the size of the sample to the risk of the required output. However, if it leads to different outputs for equal queries, the effectiveness of the SDC method is weakened because an intruder could estimate precisely the true values repeating the same queries many times. Algorithms for the POST aggregation of geographical areas (counties) has been developed for the on-farm use of agricultural chemicals on various crops for the National Agricultural Statistics Service data (Karr *et al.*, 2001). Schemes of PRE and POST source data perturbation application are shown in Figure 1.

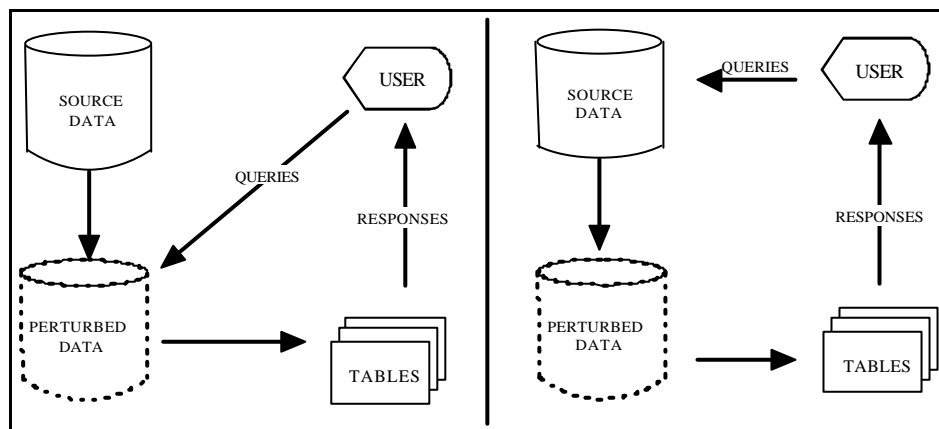


Figure 1 Scheme of SDC perturbing the source data. PRE application on the left and POST on the right.

Perturbation of the output

Perturbation of the output can be applied either PRE (for static sites) or POST (dynamic sites). The POST addition of random noises to cell values has the drawback due to repeated querying, just like the addition of random noises to the source data POST. POST cell suppression must be audited and repeated consistently; for example if a cell has been released, then it cannot be suppressed in a subsequent release, or if it is suppressed in one release, then it must be suppressed in all releases. This implies that a record of all releases must be kept. On-the-fly suppression does not seem to be practical because it requires complex and lengthy computations. Rounding cell values to a constant base does not have particular drawbacks, however it does not generally give valid protection. More effective rounding techniques, such as controlled rounding, do have the drawback of not being consistent for different tables, hence allow for disclosure using overlapping queries. To our knowledge the release from simulated data is still under

⁶ American Fact Finder: <http://factfinder.census.gov/servlet/BasicFactsServlet>

⁷ <http://sda.berkeley.edu:7502/IBGE>

study (Fienberg and Makov, 2001) but seems worth of interest, especially towards the protection of V-Labs. Schemes of PRE and POST applications of output perturbation are shown in Figure 2.

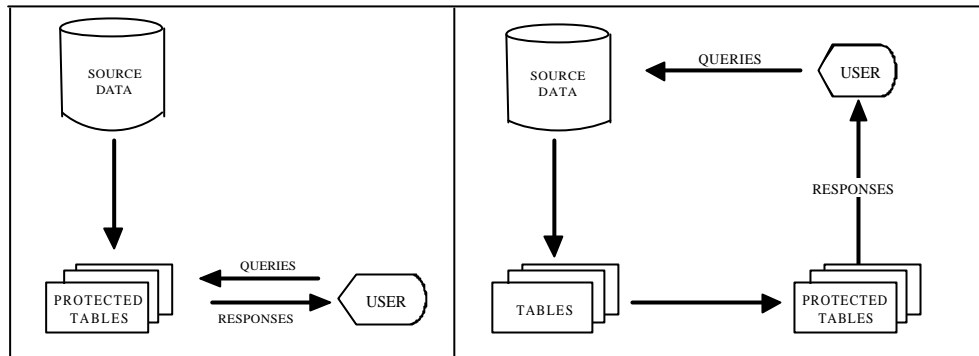


Figure 2 Scheme of SDC perturbing the output. PRE application on the left and POST application on the right.

Restriction of the query-set allowed

The restriction of the query-set allowed (Hoffman, 1977) is useful for reducing the disclosure risk connected to cumulative knowledge and linked tables. The simplest restriction gives either an approval or a denial to a query. A more elaborate type of restriction may offer a third possibility: "protect and release" or "release a simulated table" (Fienberg et al., 1998). Automatic SDC systems geared towards preventing disclosure from cumulative knowledge applying SDC sequentially in response to a series of queries were recently proposed (Keller-McNulty and Unger, 1998 and Fienberg *et al.*, 1998).

Generic restrictions can be set for all users, limiting the maximum dimensions of releasable tables or excluding tables with given combinations of spanning variables. Such restrictions should really be considered PRE SDC because certain tables are banned (i.e. suppressed completely) before queries are submitted, even though restrictions take effect only after a query is submitted. Methods for the evaluation of the disclosure risk for unreleased cells, given that some marginals have been released have not yet developed for all cases. Methods for tables of counts were developed (Buzzioli and Giusti, 1999, Dobra, 2000, and Dobra and Fienberg, 2000). These methods require intense computations and, so far, have only been implemented on prototypes. *Optimal Tabular Release* (OTR) (Karr *et al.*, 2002) is a prototype system for selecting an optimal sub-set of releasable marginal count tables. The implementation exploits heuristics for faster solution but it still seems complex and very computationally demanding even for problems of average size.

Specific restrictions are more elaborate restrictions that can also be set, for example restricting the maximum number of queries or banning the overlap of certain queries to each user. These restrictions are applicable to dynamic WSDDs and must be applied as POST SDC. Specific restrictions applied to each user require login and real-time monitoring of users. They may not be effective against coalitions of intruders, but, if enforced in a reliable way, such measures can effectively reduce the need for perturbing the data. *Table server* is a prototype of POST query restriction developed within the DG project (Karr *et al.*, 2002). Table server exploits the same procedures for the evaluation of the risk of OTR but the risk is evaluated with respect to all the information that has been released before a query is submitted. It seems important to note how such approach needs careful planning in order to avoid the WSDD to be driven to release only partial information because of peculiar requests coming from early users. Schemes of PRE and POST implementation of query restrictions are shown in Figure 3.

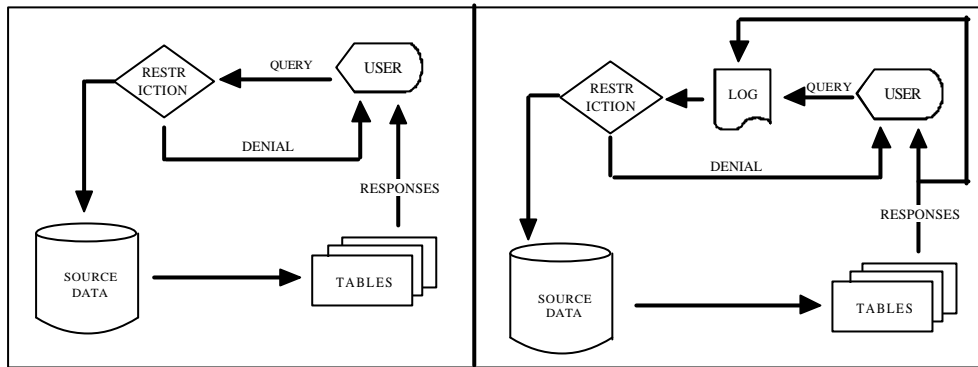


Figure 3 Scheme of queries restriction. PRE application on the left and POST application on the right.

5. Conclusions

Web dissemination will certainly become the most used way of releasing data in the future. However, the task of keeping at safe the confidentiality of respondents will unavoidably become harder as the amount of information offered on the Web increases.

The design of the methodology for the protection of a WSDD requires different choices. The first is the choice of the release policy. Depending on the type of data to be released, two opposite approaches can be followed: the first privileging the quality of the released data versus the amount of data possibly accessed; the second preferring the dissemination of an higher amount of less accurate data. The first case implies the adoption of methods that suppress few data or deny access, whereas the second requires techniques that perturb the data (source data or cell values). In some cases, it is possible to combine both strategies at the same time. The second choice consists of deciding whether to apply SDC methods PRE or POST. Certainly PRE SDC methods are more easily implemented, do not require secure access and produce safe data. The drawback is risk of over-protection and therefore lack of data access. Post SDC methods are computationally demanding but more flexible and customizable to the needs of users; however they need access restrictions and may not prevent disclosure when there is collusion among users.

Several researchers fear overprotection of Web sites and advocate a policy that broadens access to data by ensuring proper use of such data (Trivellato, 2000, Duncan et al., 2001). In order to increase the amount of data released, new strategies for applying SDC sequentially in response to a series of queries for each user, possibly through fully automated expert systems, are under study. For secure identification methods are still not yet available, a dissemination policy that distinguishes among user types might be adopted. Data products designed for public Web sites are necessarily general purpose; for this reason extreme care should be taken to avoid any possible risk of breaches of confidentiality. However, looser SDC rules could be applied to sites devoted to qualified researchers. This adaptive policy is already in use in several NSIs giving research far more opportunities, especially for microdata files.

Acknowledgements

This work was partially supported by the European Union project IST-2000-25069 CASC on “Computational Aspects of Statistical Confidentiality”.

References

- Adam, N.R. and Wortmann, J.C., 1989. “Security-control methods for statistical databases: a comparative study”. *ACM Computing Surveys*, **21**, pp. 515-556.
- Blakemore, M., 2001. “The potential and perils of remote access”. Doyle, P., Lane, J.I., Theeuwes, J.J.M. and Zayatz, L. (Eds), *Confidentiality, Disclosure and Data Access: Theory and Practical Application for Statistical Agencies*. Elsevier Science, pp. 315-337.

- Brand, R., 2002. "Microdata protection through noise addition". *Inference Control in Statistical Databases*, Lecture Notes in Artificial Intelligence. Springer-Verlag.
- Buzzigoli, L. and Giusti, A., 1999. "An algorithm to calculate the lower and upper bounds of the elements of an array given its marginals". In *Statistical Data Protection, Proceedings of the Conference*. Luxembourg, pp.131-147
- Cox, L. H., 1999 "Some remarks on research directions in statistical data protection". In *Statistical Data Protection, Proceedings of the Conference*, pp.163-176. Lisbon, Luxembourg: Eurostat.
- Dalenius, T. and Reiss, S.P., 1982. "Data-swapping: a technique for disclosure control". *Journal of Statistical Planning and Inference*, **6**, pp. 73-85.
- David, M.H., 1998. "Killing with kindness: The attack on public use data". *Proceedings of the Section on Government Statistics*, pp 3-7. (American Statistical Association)
- Dobra, A., 2000. "Measuring the Disclosure Risk in Multiway Tables with Fixed Marginals Corresponding to decomposable Loglinear Models". *Technical Report*, Department of Statistics, Carnegy Mellon University
- Dobra, A. and Fienberg, S.E., 2000. "Bounds for cell entries in contingency tables given marginal totals and decomposable graphs". *Inaugural Article, PNAS 97: 11885-11892*. Reperibile anche al sito <http://www.pnas.org/>
- Duncan, G.T. and Mukherjee, S., 2000. "Optimal disclosure limitation strategy in statistical databases: deterring tracker attacks through additive noise". *Journal of the American Statistical Association*, **95**, pp. 720-729.
- Duncan, G.T., Fienberg, S.E., Krishnan, R., Padman, R. and Roehrig, S.E., 2001. "Disclosure limitation methods and information loss for tabular data". Doyle, P., Lane, J.I., Theeuwes, J.J.M. and Zayatz, L. (Eds), *Confidentiality, Disclosure and Data Access: Theory and Practical Application for Statistical Agencies*. Elsevier Science
- Duncan, G.T., 2001. "Confidentiality and statistical disclosure limitation". In N. Smelser & P. Baltes (Eds.), *International Encyclopedia of the Social and Behavioral Sciences*. New York: Elsevier.
- Fienberg, S.E., 2000. "Confidentiality And Data Protection Through Disclosure Limitation: Evolving Principles and Technical Advances". *The Philippine Statistician* **49**, pp. 1-12.
- Fienberg, S.E., Makov E. U. and Steele, R. J., 1998. "Disclosure limitation using perturbation and related methods for categorical data". *Journal of Official Statistics*. **14**, pp. 485-502
- Fienberg, S.E. and Makov, U.E., 2001. "Uniqueness and disclosure risk: Urn models and simulation". In *ISBA 2000 proceedings*. Luxembourg:Eurostat.
- Franconi L. and Stander J., 2002. "A model based method for disclosure limitation of business microdata". *The Statistician* **51**:1, pp. 1-11
- Gouweleeuw, J.M., Kooiman, P., Willenborg, L.C.R.J. and de Wolf, P.P., 1998. "Post randomization for statistical disclosure control: theory and implementation". *Journal of Official Statistics*, **14**, pp. 463-478.
- Hundepool, A., 2001. "Computational Aspects of Statistical Confidentiality the CASC-Project". *Statistical Journal of the United Nations ECE* **18**, pp. 315-320.
- Hoffman, L. J., 1977. *Modern methods for computer security and privacy*. Prentice-Hall, Englewood Cliffs, N.J.
- Karr, A. F., Lee, J., Sanil, A. P., Hernandez, J., Karimi, S. and Litwin, K., 2001. "Web-based systems that disseminate information but protect confidentiality". In Elmagarmid, A. K. and McIver, W. M. Editors, *Advances in Digital Government*. Kluwer, Amsterdam.
- Karr, A.F., Dobra, A., Sanil, A.P. and Fienberg, S.E., 2003. "Software Systems for Tabular Data Releases". To appear in *International Journal on Uncertainty Fuzziness and Knowledge-Based Systems*. Downloadable at <http://www.niss.org/dg/index.html>
- Keller-McNulty, S. and Unger, E.A., 1998. "A database system prototype for remote access to information based on confidential data". *Journal of Official Statistics*, **14**, pp. 347-360.
- Malvestuto, F. and Moscarini, M., 1999. "An audit expert for large statistical databases". In *Statistical Data Protection, Proceedings of the Conference*, Lisbon, Luxembourg: Eurostat, pp. 29-43.
- Matloff, N. S., 1988. "Inference Control Via Query Restriction Vs. Data Modification: A Perspective". In Carl E. Landwehr (Ed.): *Database Security: Status and Prospects. Results of the IFIP WG 11.3 Initial Meeting*, Annapolis, Maryland, October 1987. North-Holland, pp.159-166.

- Trivellato,U., 2000. "Data access versus privacy: an analytical user's perspective". *Statistica*, **LX**, pp. 669-689.
- Willenborg, L. and de Waal, 2001. *Elements of Statistical Disclosure Control. Lecture Notes in Statistics*, **155**, Springer-Verlag: New-York.
- Winkler, W.E., 1998. "Re-identification methods for evaluating the confidentiality of analytically valid data". *Research in Official Statistics*, **1**, pp 87-104.
- Zayatz, L., 2002. "SDC in the 2000 U.S. decennial census". In *Inference Control in Statistical Databases*, Lecture Notes in Artificial Intelligence. Springer-Verlag.