Journal of Official Statistics
Vol. 9. No.2, 1993, pp. 511–517

Discussion
Computer Security

Bo Sundgren[1]

The problems of data security and confidentiality have always constituted an important issue in practical statistics production. Statistical offices must indeed be very careful not to forfeit the capital of confidence that they have gained vis-à-vis survey respondents. A single mistake or misjudgement could prove disastrous. On the other hand, if a statistical office is unnecessarily protective vis-à-vis its customers, the users of statistical data, it could rightfully be accused of being inefficient, uncooperative, and monopolistic. It is an extremely sensitive and challenging task to strike the right balance between the fundamental requirements on a statistical office: maximizing the productive use of statistical data in society, while preserving the confidence of information providers.

Until the late 1960s, these problems were handled by statistical offices in a heuristical way. For example rules of thumb were used for determining when certain cells in a table had to be suppressed, or otherwise manipulated, before the table was published. Release of microdata was not really a problem, since potential users outside statistical offices would seldom have the capacity to handle large volumes of data. Computers were still expensive and the mainframes of those days had a much lower capacity than the simplest and most inexpensive personal computers of today.

In the late 1960s and early 1970s a massive computerization took place in all parts of society. Researchers, planners, and decision-makers obtained access to powerful computers and developed their skills to make productive use of statistical data – in private businesses as well as in the public sector. This created pressure on statistical offices to make their data available to outside users in various forms: detailed statistical tables, produced on an *ad hoc* basis, statistical databases accessible through on-line communication, public files of anonymized microdata, etc. At the same time in many countries there was a rapidly growing concern for the possible risks of making a lot of data available in society. In particular, there was a concern for privacy matters in connection with population censuses.

In the early 1970s a lot of important research efforts were made in the area of statistical confidentiality, mainly inside the statistical offices themselves. In fact, looking back at the historical development, as represented by research documents (c.f. the list of references) it seems that a large share of important insights that we have gained concerning these problems were actually obtained during this period.

At a later stage, university researchers

[1] Statistics Sweden, S-115 81 Stockholm, Sweden.

also became interested in the problem, and a lot of money went into this research. Relatively soon, the research grew rather esoteric, mathematical, and generated intellectually stimulating problems and theoretically interesting results. Unfortunately, much of this research was essentially useless from a practical point of view. Sometimes this research must give an external observer the impression that statistical publications, files, and databases, must be leaking like riddles, leaving statistical offices with no practical alternative but closing down its activities entirely. From a theoretical point of view, disclosures always seem likely to occur, and all imaginable counter-weapons have so many negative side-effects that the battle will be lost anyhow: some medicines may be relatively efficient vis-à-vis the disease, but unfortunately they will always kill the patient as well. . . .

In contrast to this picture, created by research, the actual situation is that – to this author's best knowledge – no serious disclosure of sensitive information has ever occurred from a statistical office in a modern western society. (If such a leak had actually occurred, one could be reasonably sure that some alert newspaper would enthusiastically have scandalized the failing statistical office publicly.)

In fact, in an experiment (see Blien, Wirth, and Muller 1992 and also reported in Citteur and Willenborg 1993), where a group of scientists was asked to produce as many disclosures as possible from a number of authentical, anonymized microdata files, the result was very poor, even though the scientists were given all the resources that they might require. Of a small number of disclosures that the scientists could actually claim to have made, the majority turned out to be erroneous due to small natural errors in the underlying data.

I admit that my "summary" of some research in the area of statistical confidentiality is biased and provocative. This is intentional, because my concern is that before we spend more valuable dollars and research talents, we should seriously evaluate what we have achieved so far, and what we could possibly do to give the research a direction, which is more likely to generate constructive, practically useful results, which really come to grips with the delicate balance of interests involved in the problem.

In their article, "*Database Systems: Inferential Security*," Keller-McNulty and Unger advocate more interdisciplinary research, notably more cooperation between statisticians and computer scientists. I strongly sympathize with this view, and I would indeed go some steps further, claiming:

a. That interdisciplinary research is essential not only for *solving* problems in the area of statistical data security and confidentiality but also for identifying and describing the most relevant problems in a balanced and multi-facetted way, as well as for allocating scarce research resources where they are most urgently needed;
b. That the interdisciplinary research approach should include not only statistics and computer science but also other disciplines like systems analysis, information management, law, economics, sociology, mathematics, cryptology, etc.

One purpose of an interdisciplinary approach is to mix relevant competences from different disciplines. A more primary purpose is to get a better overview and a deeper understanding of the entire problem, making it possible to give a more nuanced description of different subproblems and how they relate to each other and the problem area as a whole. Such a well-structured problem description would

also make it possible to develop a balanced mix of tools for solving problems and subproblems in a practical situation. Furthermore, thanks to the interdisciplinary approach, each tool could incorporate relevant contributions from several disciplines.

A true interdisciplinary formulation of the problem area must start from some very general and yet very precise statements. A structured statement of the following kind could be used to span the problem space under consideration:

Somebody or something (1) for some reason (2) threatens to harm somebody (3) in some way (4) by doing something (5) with an information resource of some kind (6).

This structured statement identifies six dimensions in the problem space:

1. The threatening subject or potential intruder: a person, an organization, or "Nature," where the first two categories (persons and organizations) cause intentional threats, threats with a purpose, and the third one ("Nature") causes unintentional threats, so-called "acts of God."
2. In the case of intentional threats: the purpose of the threatening subject for performing the threatening actions. In the case of "acts of God": causal mechanisms behind the threatening events.
3. The threatened subject: a person or an organization.
4. The nature of the threat: economical damage, physical or mental damage, privacy intrusion.
5. The nature of the threatening action: destruction, manipulation, theft, unauthorized access.
6. The kind of threatened object: hardware, software, data.

When allocating research resources in the area of data security and confidentiality in statistics production, priority should be given to problems which are (i) important and (ii) specific for statistics production. Problems which are important but not specific for statistics production should be handled in a more general context. Most unintentional threats (according to the classification above) belong to this category, for example, threats due to technical malfunctions in hardware and software as well as threats due to natural disasters.

Two major categories of threats, which are important and specific for statistics production, can be defined in the following way, using the six classification categories given above.

Category 1 Threats. The threatening subject is a journalist or somebody else, who – for some political reason – wants to discredit the statistics production process or the society of which the statistics production process is a subsystem. This type of threat could focus on sensitive information about individual persons, since the disclosure of such information often has a particularly large effect on public opinion. Thus the threatened subject is (in addition to the statistical office and the society concerned) typically a person, the nature of the threat is privacy intrusion, and the purpose is political. The nature of the threatening action could be theft or some other form of unauthorized accessing of data, but – as long as the threatening subject wants to maintain some basic respectability within the society concerned – the threatening action is more likely to be based on some form of backwards identification (Block and Olsson 1976) on the basis of threatened objects such as legally available statistical information (published statistics or public files of anonymized microdata).

Category 2 Threats. The threatening subject is a person or a company who wants to

gain economical advantage by getting (exclusive) access to information about some other market actors, for example, competitors or actors with a generally large effect on the market. In this type of threats both the threatening subjects and the threatened subjects are economical actors. Indirectly the threats could be harmful to the statistics producer (almost certainly) and (possibly) to the public trust in the market economy or society at large. Thus the nature of the threat is primarily economical and secondarily political.

The two major categories of threats are completely different in many respects, and they must be met by different mixes of counter-actions. Category 2 Threats are economically rational in the sense that they are – at least in principle – based upon ordinary business considerations. Thus cost/benefit analyses are relevant for determining (a) the "size" of the threat; and, consequently, (b) the necessary "size" of counter-actions. The amount of resources, which the threatening subject is ready to spend on threatening actions should not exceed the value of the benefits obtainable through these actions.

Economical rationality in the sense just mentioned is not really applicable to Category 1 Threats. On the other hand, most potential threatening subjects behind this category of threats can probably be assumed to be relatively law-obedient. If they do not subscribe to the basic principles of a democratic society regulated by laws, they could easily find much more efficient ways of terrorizing society and harrassing individual citizens than by misusing statistical data.

Since Category 2 Threats are typically based upon some (explicit or implicit) cost/benefit analysis, the design of an appropriate scheme of protection measures could also be based upon some kind of cost/benefit analysis. The protection measures should be so expensive to circumvent, that the potential intruder does not even find it worthwhile to try.

A special problem for an important subset of Category 2 Threats is where protection is not possible at any cost, due to some typical characteristics of the economical actors in a modern society. In most small or medium-sized countries there are a small number of economical actors, who completely dominate the activities within a certain sector of society. For example, in Sweden there are two car manufacturers. Moreover, whatever these few actors do will significantly affect the economy of the whole country, so the information cannot be suppressed or distorted, if the statistics producer wants to give the users of statistical information a meaningful picture of the economy of the country. The only feasible ways out of this dilemma seem to be:

either to base statistics on public data only;
or

to make agreements with the respondents that they do not mind that their data can be derived or estimated with relatively high precision from published statistical data.

If we turn to Category 1 Threats, conventional, "money-oriented" cost/benefit analyses are not applicable. However, we can apply a generalized concept of cost/benefit analysis, where the intruder's benefit is the gain in political sympathy that the intruder could obtain by discrediting the statistics producer on the basis of disclosures of sensitive information about individual respondents on the basis of published statistics or publicly disseminated statistical microdata. Under normal circumstances such a privacy intrusion would hardly be appreciated by the general public, if it were

based upon criminal behaviour. Thus it seems that legislation and legal contracts would be appropriate protection measures against this category of threats.

In Sweden, since the beginning of 1993, a new law is in force – the law of official statistics – which criminalizes backwards identification of individual objects on the basis of official statistics (in combination with other information). The crime is called "unlawful identification" and can be punished with one year's imprisonment, unless it could be regarded as an even more serious crime according to some other law. The only Category 1 Threat that is not protected against by this law seems to be a situation where the disclosure of sensitive information about an identified individual object is so obvious that no deliberate effort is necessary to accomplish the disclosure. In most practical situations it should be a relatively simple and straightforward task to protect against such obvious disclosures.

I think that this very brief analysis, although simplified, clearly demonstrates some important points, which may also be regarded as conclusions of my reflections as well as a starting-point for future, interdisciplinary research efforts:

1. There are a relatively small number of characteristical types of threats to a statistics producer which are important from a practical point of view, and which are specific for statistics production.
2. The different types of threats, exemplified above by Category 1 Threats and Category 2 Threats, show different characteristics between themselves, each type of threat calling for a characteristic mix of protection measures and counter-actions.
3. For most kinds of threats, which are specific for statistics production, technical considerations – including mathematical/statistical arguments so amply represented in the literature – should indeed have a subordinate role.

To prevent misunderstandings, I should mention that the list of references below contains mainly two categories of articles, both of which I regard as extremely valuable, also from the point of view of practical statistics producers:

1. Review articles, giving well-structured overviews of entire problem areas.
2. Articles representing "intellectual breakthroughs" giving important insights for a good understanding of a certain problem area.

To conclude, where do we stand after 25 years of research in the problem area of data security and confidentiality in statistics production, and where do we go from here?

For the time being, I think that research has produced enough knowledge about a multitude of intellectually challenging, more or less relevant subproblems within the problem area. The subproblems have often been narrowly defined, so as to make them more suitable for study by means of formal theories and methods from the tool-boxes of mathematics, statistics, and computer science. Despite a number of useful review articles, covering somewhat larger parts of the problem area, sometimes in an interdisciplinary way, we are still waiting for more whole-hearted, constructive attempts to organize the large volumes of important but rather piecemeal knowledge fragments that we have so far collected into a well-structured, comprehensive, and – most importantly – practically useful body of knowledge (or theory) for the problem area as a whole. As I have indicated in this article, such a knowledge organization effort could be based upon a struc-

tured, multidimensional description of the entire problem space. It should identify the cells in this problem space, which are relevant and specific for official statistics production, and it should give guidelines for the design and practical implementation of a well orchestrated system of methodological tools and organizational actions, which an official statistics producer could actually use to protect the confidence of its information providers while maximizing the productive use of statistical data in society. This is a challenge, which requires (c.f. Mason, McKenney, and Copeland 1991) visionary, "business-oriented" leaders as well as system and organization oriented "maestroes" in addition to talented "supertechs" from more formally and technologically oriented disciplines like mathematics/statistics and computer science. I am afraid that with respect to this challenge, research has not moved us very far from where we were 25 years ago. Who picks up the gauntlet?

References Cited in the Text

Blien, U., Wirth, H., and Muller, M. (1992). Disclosure Risk for Microdata Stemming from Official Statistics. Statistica Neerlandica, 46, 69–82.

Block, H. and Olsson. L. (1976). Backwards Identification of Person Information. Statistical Review, 14, 135–144.

Citteur, C.A.W. and Willenborg, L.C.R.J. (1993). Public Use of Microdata Files: Current Practices at National Statistical Bureaus. To appear in the Journal of Official Statistics.

Mason, R.O., McKenney, J.L., and Copeland, D.G. (1991). Developing an Historical Tradition in MIS Research. The Harvard MIS History Project.

References Not Cited in the Text

Adam, N.R. and Wortmann, J.C. (1989). Security-Control Methods for Statistical Databases: A Comparative Study. ACM Computing Surveys, 21, 515–556.

Bethlehem, J.G., Keller, W.J., and Pannekoek, J. (1990). Disclosure Control of Microdata. Journal of the American Statistical Association, 85, 38–45.

Biggeri, L. and Zannella, F. (1991). Release of Microdata and Statistical Disclosure Control in the New National Statistical System of Italy: Main Problems, Some Technical Solutions, Experiments. Conference of the International Statistical Institute, Cairo.

Cassel, C.M. (1974). On Probability Based Disclosures in Frequency Tables. Statistics Sweden, R&D Reports.

Cox, L.H. (1980). Suppression Methodology and Statistical Disclosure Control. Journal of the American Statistical Association, 75, 377–385.

Chin, F.Y. and Özsoyoglu, G. (1982). Auditing and Inference Control in Statistical Databases. IEEE Transactions of Software Engineering SE-8, 6, 574–582.

The Committee on Federal Agency Evaluation Research (1975). Protecting Individual Privacy in Evaluation Research. Washington, D.C.: National Academy of Sciences.

Dalenius, T. (1977). Towards a Methodology for Statistical Disclosure Control. Statistical Review, 15, 429–444.

Dalenius, T. (1988). Controlling Invasion of Privacy in Surveys. Statistics Sweden, R&D Reports.

Denning, D.E. (1978). A Review of Research on Statistical Database Security. New York: Academic Press.

Denning, D.E., Denning, P.J., and Schwartz, M.D. (1979). The Tracker: A

Threat to Statistical Database Security. ACM Transactions on Database Systems, 4, 76–96.

Fellegi, I.P. (1972). On the Question of Statistical Confidentiality. Journal of the American Statistical Association, 67, 7–18.

Fellegi, I.P., and Phillips, J.L. (1974). Statistical Confidentiality: Some Theory and Applications to Data Dissemination. Annals of Economic and Social Measurement, 3, 399–409.

Flaherty, D.H. (1989). Protecting Privacy in Surveillance Societies: The Federal Republic of Germany, Sweden, France, Canada, and the United States. University of North Carolina Press, Chapel Hill.

Frank, O. (1973). Reconstruction of Individual Data from Classificational Frequency Distributions. University of Uppsala.

Hoffman, L.J. and Miller, W.F. (1970). Getting a Personal Dossier from a Statistical Data Bank. Datamation, 16, 74–75.

Michalewicz, Z. (1981). Compromisability of a Statistical Database. Information Systems, 6, 301–304.

Nargundkar, M.S. and Saveland, W. (1972). Random Rounding: A Means of Preventing Disclosure of Information about Individual Respondents in Aggregate Data. Proceedings of the American Statistical Association.

Olsson, L. (1973). Measures to Protect Privacy in a Statistical Data Base. Statistics Sweden, R&D Reports.

Rapaport, E. and Sundgren, B. (1975). Output Protection in Statistical Data Bases. Conference of the International Statistical Institute, Warsaw.

Richardson, M. (1972). National Statistical Information Systems: Background Note on Confidentiality Processes. UN Computing Research Centre, Bratislava.

Schlörer, J. (1975). Identification and Retrieval of Personal Records from a Statistical Data Bank. Methods Inform Med, 14, 7–13.

Schlörer, J. (1976). Confidentiality of Statistical Records: A Threat Monitoring Scheme of On-Line Dialogue. Methods Inform Med, 15, 36–42.

Sundgren, B. (1972). Security and Privacy of Statistical Data Bases. Statistical Review, 10, 299–312.