INTERNATIONALES SEMINAR ÜBER DIE STATISTISCHE GEHEIMHALTUNG

PROTOKOLLE

INTERNATIONAL SEMINAR ON STATISTICAL CONFIDENTIALITY

PROCEEDINGS

SÉMINAIRE INTERNATIONAL SUR LA CONFIDENTIALITÉ STATISTIQUE

ACTES

8-10 September 1992
Dublin, Ireland

Themenkreis / Theme / Thème
Verschiedenes / Miscellaneous / Divers
Reihe / Series / Série
Methoden / Methods / Méthodes

9
E

STATISTISCHES DOKUMENT ☐ STATISTICAL DOCUMENT ☐ DOCUMENT STATISTIQUE

Auf Recyclingpapier gedruckt ☐ Printed on recycled paper ☐ Imprimé sur papier recyclé

TECHNOLOGICAL ASPECTS OF CONFIDENTIALITY: NEW TECHNOLOGY - THREAT OR GREATER PROTECTION?

Lars Thygesen
Danmarks Statistik

1. INTRODUCTION

For more than a hundred years statisticians have been aware of the problems of confidentiality. Before the start of the computer age, the problems were rather easy to handle, at least in principle: Tables could be published, and the original forms with personal information must be kept strictly secret. This called for a good safe and for rules laying down how to treat paper forms.

With the advent of (mainframe) computers in statistics production, these problems became more complicated. Now large amounts of individual data could be stored on relatively small media, notably magnetic tapes. These tapes would be handled in a central 'computer-department' and would be kept under strict control. The tapes could be easily copied and processed, if you possessed the necessary equipment. But fortunately, only professional organisations possessed such equipment.

Until recently it has not been practicable to process great amounts of individual data for statistical purposes on micro computers. In central statistical offices (CSOs) we have established rules and procedures governing our own use of the data as well as the ways to release data for research purposes.

Now things have changed with the advent of powerful and inexpensive micro computers, local area networks, and new telecommunications facilities. We have not yet fully realized what this is going to imply for our procedures.

2. THE MICRO REVOLUTION: INSIDE THE CSO

It is well known that the spread of micro computers has by now removed the data processing monopoly from the big computing centres. The micros have entered the desks of researchers and administrators. And in our homes, we see that every nursery school child commands enormous computing power. This has changed society, and it has certainly changed the working conditions of CSO's.

In the first place, it is now advantageous to carry out at least part of the basic production of official statistics on the micros in the CSO. Consequently there is a strong incentive to store some personal data on micros for processing, rather than on a central mainframe that can be protected by special measures: Data are moving from protected areas into ordinary office areas, and data are stored on PCs with poor security characteristics, in stead of being stored on computers that can support very strong security systems. It is difficult, if not impossible, to control the pressure for this change

The horror scenario is that somebody walks into or breaks into an office in the CSO and steals the micro with its data. At home the thief can more easily, in peace and quiet, break the rather weak security systems that are available for the micros. The intruder might be somebody who wishes to use the data for personal gain or, perhaps more likely, somebody who wishes to expose the vulnerability or incompetence of Government statistics. In either case we have a disaster.

How should data be protected against this risk? One possible answer would be to introduce strict access control or other rules on behaviour in the office areas, but it seems unlikely that this would solve the problem. It is impossible to maintain the same level of security in normal offices as in computer centres.

The only appropriate answer we have been able to come up with in Danmarks Statistik, is to forbid the storing of confidential data on micros. In practice, this has led us to introduce a computer architecture with all of the micros being net-stations without hard-disks and floppy disks, tied together in a local area network (LAN). Whenever data are going to be processed, they have to be fetched from a server that is physically placed in the computing centre under special protection.

It should be added that this way of organising things creates some problems, because it adds to the traffic on the LAN and thus requires a high capacity network, and at the same time the PCs loose their independence: They will only be able to run as long as the network is in operation, and this means that breakdowns must not occur. On the other hand, this architecture also has other merits: It eases common backup procedures and makes it possible to relieve the many network users of other tedious tasks, such as updating the system with new facilities and software versions. And from the view-point of the company, it is important that backup of corporate data is actually carried out, independent of whether employees forget to do it.

3. THE MICRO REVOLUTION: THE USERS OF STATISTICS

What goes on in the statistics users' world is also extremely important to the question of confidentiality, as seen from the point of view of official statistics. And here it is evident that the Micro Revolution has had a very strong effect.

A few years ago, users were rarely in a position to carry out heavy data processing tasks, like aggregating millions of records, or making multi-dimensional categorical analysis. This demanded mainframes, and they were heavy and expensive.

The micros are now available at the desk top of every user of statistics. And the users want to make use of their processing power. Often they want to take over as large a proportion of the processing as possible, which is quite sensible as the micros have already been bought, so the processing power is 'free'. Especially when you carry out multi-variate analysis on socio-demographic data, the process is helped very much if you have access to micro data so alternative models can easily be tested.

Thus, the demand pressure for data, especially micro data, has grown. At the same time, the risk of disclosure connected with releasing such data has risen even more because of the large amount of idle data processing power that is spread over society. This power can be used for trying to break anonymity, that is to disclose the identity of persons or firms from detailed statistical material. This combination really creates a new threat to confidentiality.

Some authors (e. g. Marsh et al. 1991) have argued that it is possible to avoid this danger by giving access only to 'safe data', that is fully anonymised microdata which do not contain unique combinations of 'key variables' (or contain a very low fraction of such uniques). This is a practice that has been used in some countries for many years. But in my opinion this will not solve the problems. Either the data will lack the necessary detail and will thus not satisfy researchers' needs; or it would still be feasible to break the anonymity by combining, outside the CSO, a number of different data sets, each of them 'safe'. This risk is discussed in Fellegi 1972. Since then, the problems have obviously multiplied because of the growth in data supply, in data demands, and in data processing capability. The idea that disclosure can be controlled if a large

number of microdata sets are disseminated must be discarded.

According to Danish experience, social researchers very often will not be satisfied with access to 'safe data', because their data needs will be extremely specific. They will want to combine data from a broad spectre of socio-demographic statistics, and more and more frequently, they will want longitudinal data in order to test specific hypotheses. Data like that can in many cases be produced by applying matching techniques. It is a shame if we have to deny access to these excellent research opportunities because of confidentiality.

So we need to be able to offer other possibilities than access to safe data with insufficient contents.

4. SAFE SETTINGS?

If statistics users are to be satisfied and confidentiality is to be respected at the same time, I see no alternative to keeping the very detailed statistical data inside the CSO. Each project needing such data must be evaluated with regard to the risk of disclosure. Depending on the circumstances, we should be able to offer a number of different access methods, providing 'safe settings' (Marsh et al. 1991). I am going to discuss some methods being used in Denmark, one of them an innovation being tested at the moment.

Firstly, the old-fashioned way of dealing with the problem is that the CSO refuses to give physical access to the data, and offers to carry out the processing on behalf of the researcher. In this way the researcher will only receive non-confidential data. The draw-back of this method is that it is a heavy procedure because the researcher has to explain his intentions to somebody else, who will carry out the job. This may cause misunderstandings and errors. It may also be expensive.

Secondly, we have an arrangement called *the Inhouse Researcher*. In this case, the researcher may process a data set that is not safe, provided the work is carried out at the premises of the CSO. The researcher has to sign an agreement, making his legal status similar to that of a sworn employee at the CSO. The confidential data rests within the computer system of the CSO. The data set is created especially for each research project and contains only the anonymised personal data necessary for that project. This arrangement has been welcomed by many researchers, while others find it unsatisfactory, because it is inconvenient to have to work in the CSO (it should be noted that the CSO in Denmark is situated in only one location which makes it necessary for some researchers to travel a long way).

Trying to respond to discontent with the Inhouse Researchers arrangement, Danmarks Statistik has developed a new service: *The Researchers´ Mailbox.* The idea is that a special data set with detailed but anonymous data is created, matching the needs of the individual researcher. This data set is kept at the host computer in the CSO. A micro in Danmarks Statistik is set up with an electronic mailbox, ready to receive statistical applications programs prepared by the researcher. The mailbox is emptied twice a day and the researchers' jobs are transferred by CSO officials to the host computer where they are run. The result (in electronic form) is inspected and - if accepted - transferred to the mailbox. The researcher may then fetch his output from the mailbox via the public data network.

The concept of the Researchers' Mailbox is being tested right now, using one actual research project as a test case. It seems to be working well and the arrangement poses no special confidence control problems. The problems concern the practical procedures of handling incoming jobs from researchers and exercising control of the results, without undue inconvenience to neither the researchers nor the CSO staff.

5. DATA COMMUNICATION AND CONFIDENTIALITY

The most promising development in Information Technology at present seems to be the advance of data telecommunication facilities. Even though this development has already had a great impact in the past few years, we have just seen the beginning.

5.1. Local Area Networks

I have already mentioned the advance of *Local Area Networks (LANs)*, making it easy within an organisation to share data (and other resources) and send messages between workstations. When the data that are sent consist of confidential statistical information, it becomes very important that you can control the communication. You need to be absolutely sure about the identity of the person in the other end of the line. And it must be secured that nobody else can listen on the line.

At present, security on LANs is often rather vulnerable. The users' data can normally be accessed with a personal access code (password). If this password is stolen or guessed by somebody else, the access rights of the owner can be misused.

5.2. Long distance communication

Electronic data communication over longer distance is also spreading. Until now, the common way of sending very detailed information (when it is admissible to send it) has been by hand of messenger or by classified mail, the typical media being tapes or diskettes. Now, we see a growing pressure to switch to online connection. In private business, Electronic Data Interchange (EDI) is growing in importance as the way in which companies exchange (confidential) business information. When the companies are to deliver basic data to CSOs for statistical purposes, they want to use the same effective way of communicating. And increasingly, statistical offices want to use EDI techniques also when they exchange statistical information between themselves or with users of statistics. In Eurostat and in the UN Economic Commission for Europe, a lot of work is being done preparing an international EDI standard for statistical messages (Lebaube 1991, Maurer et al. 1992).

In addition it should be mentioned that statistical online services have already been established in some countries, giving users immediate access to large statistical databases (Gewalli 1989). In my opinion, it will be the natural way of seeking statistical information for tomorrow's users of statistics.

In summary, it is a reasonable guess that in a few years, a large proportion of the basic data going into CSOs, as well as of the statistical output to the users, will be transmitted online. What will be the implications to the protection of confidentiality?

Some of the computing facilities of the CSO obviously will have to be connected with some external network. This could be a public network or a private network of leased lines. In either case, we will have the problem of identifying exactly the person or the computer in the other end of the line. We have all read shocking stories in the newspapers on hackers finding their way into computers supposed to be very secure. None of these stories so far seems to relate to statistics, but there is no reason why this might not happen.

5.3. New ways of organising statistical work

In the near future, the growing use of powerful portable micros will lead to new ways of organising

more threats are on their way. At the same time, new technologies may be used to create new ways of protection against those same threats. It seems to me that if we make careful use of the opportunities, we should be able to keep a high level of protection, and to maintain public confidence in official statistics.

REFERENCES

Biggeri, L. and Zanella, F. (1991): Release of Microdata and Statistical Disclosure Control in the New National Statistical System of Italy: MainProblems, Some Technical Solutions, Experiments. Bulletin of the International Statistical Institute, Vol. LIV, Book 1. Cairo.

Fellegi, I. P. (1972): On the question of Statistical Confidentiality. Journal of the American Statistical Association, Vol. 67, No. 337. Washington D. C.

Fellegi, I. P. (1991): Editorial: Maintaining Public Confidence in Official Statistics. Journal of the Royal Statistical Society, Series A, 154, 1. London.

Gewalli, L. E. (1989): Statistical Data Bank Services in Denmark. Statistical Journal of the UN Economic Commission for Europe, Vol. 6, No. 1. Geneva.

Habermann, H. and Weiss, P. (1992): The future of EDI in the United States Statistical System. Proceedings of Conference on New Techniques and Technologies for Statistics (Preprint). Bonn.

ISO 7498-2 (1989): Information Processing Systems - Open System Inter connection Reference Model - Security Architecture. Geneva.

ISO 8731-1 (1987): Banking - Approved Algorithms for Message Authentication - Data Encryption Algorithm. Geneva.

ISO/IEC 9798-1: 1991(E) (1991): Information Technology - Security Techniques - Entity Authentication Mechanisms. Geneva.

Lebaube, P. M. (1991): EDI et statistique. Un enjeu pour les statisticiens. Bulletin of the International Statistical Institute, Vol. LIV, Book 3. Cairo.

Malamud, C. (1991): STACKS. Interoperability in Today's Computer Networks. Prentice Hall, N. J. Marsh, C., Dale A. and Skinner, C. (1991): Safe Data versus Safe Settings: Access to Customised Results from the British Census. Bulletin of the International Statistical Institute, Vol. LIV, Book 1. Cairo.

Maurer, A. and Kubler, J. E. (1992): EDI and Statistics - The Need for a Generic Message. Proceedings of Conference on New Techniques and Technologies for Statistics (Preprint). Bonn.

Spieker, F.(1992): The balance between Demands for Statistics and Statistical Confidentiality in Denmark. International Seminar on Statistical Confidentiality. Dublin.

Thygesen, L. (1986): Data Protection in a National Register-Based Statistical System. In 'Protection of Privacy, Automatic Data Processing and Progress in Statistical Documentation', Eurostat News, Special edition, Theme 9, Series C. Luxembourg.