

Topic (ii): software and computing developments

CRYPTOGRAPHIC TECHNIQUES IN STATISTICAL DATA PROTECTION

Submitted by Universitat Rovira I Virgili, Spain¹

Contributed paper

Summary

1. Statistical data protection is a broad concept that encompasses all aspects of the protection of sensitive statistical data. Statistical disclosure control is a part of statistical data protection whose mission is to thwart inference of sensitive data from published statistical information. However, there are several security/functionality holes in the normal operation of a national statistical institute that cannot be walled up by statistical disclosure control techniques. This paper identifies some of these problems and sketches cryptographic solutions for them.

KEYWORDS: Statistical data protection, Statistical disclosure control, Cryptography, Official statistics.

I. Introduction

2. With the exponential growth and importance of computer networks, storing, distributing and handling sensitive statistical data over the Internet is becoming a key issue of concern for National Statistical Institutes (NSIs). Encryption techniques can provide invaluable help to overcome these problems. This paper will concentrate on two shortcomings in the area of statistical data protection that are not being covered by statistical disclosure control. Specifically:

- a) the need to provide the general public with data which are disclosure-protected but which are still usable to obtain exact statistics.
- b) the need to store, distribute and handle confidential statistical data (e.g. raw respondents' data) in a secure way.

Cryptographic solutions to deal with the above problems will be outlined in Sections 1 and 2, respectively.

¹ Prepared by Josep Domingo-Ferrer.

II. Obtaining Exact Statistics from Disclosure-Protected Data

3. Statistical disclosure control attempts to keep individual information anonymous when releasing macrodata (contingency tables) and microdata (individual records). Such data can be released by publishing printed reports or following a set of queries to a statistical database.

4. Disclosure control methods (Willenborg and De Waal, 1996) rely on one or a combination of the following basic operations: random perturbation, data suppression (or query set size control for on-line statistical databases) and random sampling. Therefore, disclosure control methods yield data only usable for consultation and APPROXIMATE computation at an unclassified level.

5. In Domingo-Ferrer (1996), a cryptographic approach was shown allowing a classified level (National Statistical Institute) to take advantage of unclassified (e.g. subcontracted) computation on disclosure-protected data. The same approach allows the NSI to offer to the general public the possibility of obtaining EXACT statistical results from disclosure-protected data. The idea is to extract, with little classified effort, exact statistics from statistics computed on disclosure-protected data at an unclassified level. Ad hoc encryption transformations (Blakley and Meadows, 1985) or privacy homomorphisms (encryption transformations allowing a restricted set of operations to be carried out directly on encrypted data; Rivest, Adleman and Dertouzos, 1978) are the cryptographic tools that can support that approach. Further information on a prototype that implements the above ideas can be found in Domingo-Ferrer and Sanchez del Castillo (1998) and Domingo-Ferrer, Sanchez del Castillo and Castilla (1997).

III. Secure Handling, Distribution and Storage of Confidential Statistical Data

6. Raw respondents data or otherwise unprotected confidential data should be stored, distributed and handled securely by NSIs. Encryption is an obvious solution for this problem and is already explicitly mentioned by recent statistical laws (e.g. Draft Statistical Law of Catalonia, 1998).

7. Secure storage can be achieved by keeping confidential data files encrypted. Particular encryption transformations can be chosen to minimize the storage space needed, to maximize encryption/decryption speed or even to allow some operations to be performed directly on encrypted data (see Section 1).

8. Secure distribution of confidential data normally requires all communicating parties to have cryptographic facilities and certified public encryption keys. In Polemi and Kokolakis (1998), a solution for the connection of NSIs with each other and with the outside world is outlined.

9. Secure handling of raw respondents data currently "relies" on legal non-disclosure agreements signed by the data collectors, who are very often people temporarily hired for a given survey. This obviously precarious legal security measure should be complemented/replaced by technical measures. A possibility that becomes realistic with the current state of technology would be the following protocol.

- a) The data collector handles a laptop to the respondent.
- b) The respondent enters his/her answers to the questions of the survey.
- c) The answers are encrypted by the laptop using the public encryption key of the NSI conducting the survey. This public key can be prerecorded in the laptop or can even be published in the newspaper and typed by the respondent in real time.

10. A variation of the above protocol could allow the respondent to answer a survey without the physical presence of the data collector. In that scenario, the respondent could use his/her own home computer to supply his/her answers; the answers would be encrypted by the home computer and then sent to the NSI via Internet. Some randomization would probably be needed to prevent a wiretapper from identifying the clear responses from the encrypted responses.
11. The above-mentioned cryptographic solutions will be discussed in more detail in the full paper.

References

- G. R. Blakley and C. Meadows (1985)
 "A database encryption scheme which allows the computation of statistics using encrypted data", in Proceedings of the IEEE Symposium on Research in Security and Privacy, New York: IEEE CS Press, pp. 116-122.
- J. Domingo-Ferrer (1996)
 "Privacy homomorphisms for subcontracting statistical computation", in Proceedings of the 3rd International Seminar on Statistical Confidentiality, Ljubljana: Eurostat-Statistical Office of the Republic of Slovenia, pp. 189-191.
- J. Domingo-Ferrer and R. X. Sanchez del Castillo (1997)
 "An implementable scheme for secure delegation of computing and data", in Information Security-ICICS'97 (Lecture Notes in Computer Science 1334), eds. Y. Han, T. Okamoto and S. Qing, Berlin: Springer-Verlag, pp. 445-451.
- J. Domingo-Ferrer, R. X. Sanchez del Castillo and J. Castilla (1998)
 "Dike: A prototype for secure delegation of statistical data", in Proceedings of SDP'98, ed. J. Domingo-Ferrer, Amsterdam: IOS Press (to appear).
- D. Polemi and G. Kokolakis (1998)
 "A secure network of European Statistical Offices over the Internet", in Proceedings of SDP'98, ed. J. Domingo-Ferrer, Amsterdam: IOS Press (to appear).
- R. L. Rivest, L. Adleman and M. L. Dertouzos (1978)
 "On data banks and privacy homomorphisms", in Foundations of Secure Computation, eds. R. A. DeMillo et al, New York: Academic Press, pp. 169-179.
- L. Willenborg and T. de Waal (1996)
 Statistical Disclosure Control in Practice, New York: Springer-Verlag.