

Topic (ii): software and computing developments

## **CRYPTOGRAPHIC TECHNIQUES IN STATISTICAL DATA PROTECTION**

Submitted by Universitat Rovira I Virgili, Spain<sup>1</sup>

### **Contributed paper**

#### **I. INTRODUCTION**

1. The production of official statistics can be regarded as a process with three main steps: *data collection*, *data processing* and *data dissemination*. New information technologies have a twofold impact on the above steps:

- Excluding security problems, electronic information is clearly much more convenient to collect, process and disseminate than paper-based information.
- When security is considered, unprotected electronic information can be searched, copied, counterfeited or deleted by intruders much more easily than unprotected paper-based information.

2. The above remarks explain that an increasing amount of research in official statistics is devoted to *protecting* information, so that the advantages of using new technologies can be enjoyed without jeopardizing statistical confidentiality. In Buzzigoli and Giusti (1998), a distinction is made between two basic concepts for statistical confidentiality:

**SDC:** Statistical Disclosure Control (SDC) techniques can be defined as the set of methods to reduce the risk of disclosing information on individuals, businesses or other organizations. Such methods are only related to the dissemination step and are usually based on restricting the amount of information released (see Willenborg and De Waal (1996) for more detail).

**SDP:** Statistical Data Protection (SDP) is a more general concept which takes into account all three steps of production. SDP is multidisciplinary and draws on computer science (data security), statistics and operations research.

3. Cryptography, the science that deals with the design of encryption or cipher systems, can help solving several problems in SDP that are not addressed by SDC. This paper will elaborate on cryptographic solutions for SDP problems. Section II deals with problems related to the data collection

---

<sup>1</sup> Prepared by Josep Domingo-Ferrer and Josep M. Mateo-Sanz (Universitat Rovira I Virgili) and Ricardo X. Sánchez del Castillo (University of Barcelona).

step. Issues connected with the data processing step are addressed in Section III. Section IV has to do with data dissemination. Section V is a conclusion.

## II. CRYPTOGRAPHY AND DATA COLLECTION

4. Traditionally, secure handling of raw respondent data has been dependent on legal non-disclosure agreements signed by the data collectors, who are very often people temporarily hired for a given survey. The fact that raw respondent data can *actually* be seen by the data collector (or even worse by an intruder) is the most prominent security problem in traditional data collection. Legal security measures alone do not suffice. Therefore, the following are desirable objectives:

- To prevent the data collector (and of course any unauthorized person) from seeing clear respondent data. See Subsection A.
- To thoroughly suppress the role of the data collector by carrying out data collection remotely over the Internet. See Subsection B.

### A. Hiding data from the collector

5. A possibility that becomes realistic with the current state of technology would be the following protocol:

#### Protocol 1

1. *The data collector handles a laptop to the respondent*
2. *The respondent enters his/her answers to the questions of the survey.*
3. *The answers are encrypted by the laptop using the public encryption key of the NSI conducting the survey, so that only the NSI will be able to decrypt them using its corresponding private key (see Note 1). The NSI public key can be prerecorded in the laptop or can even be published in the newspaper and typed by the respondent in real time.*

**Note 1.** In a public-key cryptosystem (Diffie and Hellman 1976), each user  $u$  has a key pair  $(PK_u, SK_u)$ , where the public key  $PK_u$  is publicly known and the private key  $SK_u$  is only known to  $u$ . A message encrypted under the public key can only be correctly decrypted under the private key. In this way, everybody is able to send secret messages to  $u$ , since  $PK_u$  is public and  $SK_u$  is only known to  $u$ . Normally, some kind of certification is assumed to guarantee that  $PK_u$  is *really*  $u$ 's public key. RSA (Rivest, Shamir and Adleman 1978) is the best known public-key cryptosystem.

### B. Suppressing the collector

6. An alternative way to eliminate the security risks associated to the data collector is to suppress its role completely. In fact, the respondent can 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 (using public-key cryptography as described in Protocol 1) 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.

## III. CRYPTOGRAPHY AND DATA PROCESSING

7. Since NSIs are committed to statistical confidentiality, unprotected confidential data are

assumed to be processed, transferred and stored in a secure environment. This poses several problems:

- How should unprotected respondent data be stored? How to transfer non-disclosure-protected data between the various locations of a NSI? Is it possible to use an open network such as the Internet to interconnect these locations?
- Does the NSI commitment to statistical confidentiality preclude the use of external untrusted subcontractors to perform computations on confidential data?
- In the "data-shop" context, how to handle requests by customers who wish to have some statistical computations done on a confidential data set? Since unprotected data cannot be exported and disclosure-protected data do not yield exact statistics, should computation always be carried out by the NSI?

Encryption can be very helpful for solving the above problems and is already explicitly mentioned by recent statistical laws (*e.g.* Draft Statistical Law of Catalonia, 1998).

#### **A. Storage and transfer of non-disclosure-protected data**

8. 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 Subsection B). 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 sketched.

#### **B. Delegation of computing and data**

9. The need for delegating statistical data arises when the data owner (*e.g.* NSI) must have its data handled by an untrusted external party, who can be either a customer or a subcontractor. The following are two practical applications:

**Example 1.** A computing delegation problem appears whenever a (small) company wants to use external computing facilities to do some calculations on corporate confidential data. A very common variant of this situation is a medical research team using a (insecure) university mainframe for processing confidential healthcare records. The reason for using external facilities may be the complexity of the calculations but also the huge size of the data set. ♦

**Example 2.** Data delegation problems appear in the interaction between public administrations at several levels. For example, municipalities cooperate with NSIs in statistical data collection. In return, municipalities would like to be able to analyze the whole collected data set (pooled from all municipalities). But only NSIs are usually authorized to hold nation-wide individual census data. A similar problem occurs in any federal-like structure (European Union, U.S.A., Germany, etc.). Member states cooperate with federal agencies in collecting data from individuals, companies, etc. In return, states would like to be able to analyze data at a federal level. A secure solution in both scenarios above is for the organization owning the whole data set to perform (probably for free) the analyses requested by the cooperating organizations. But then the data owning organization becomes a bottleneck and is forced to waste time and resources in uninteresting tasks. A better solution would be for the data owner to delegate data in a secure way and reduce its role in subsequent analyses to a minimum. ♦

10. A secure solution to delegation must satisfy two basic requirements:

- *Data secrecy.* The data owner does not want the data handler (the untrusted party performing the computations) to learn the confidential data being processed.
- *Computation verifiability.* The data owner wants to make sure that the computations performed by the data handler are correct. Note that, in computing delegation, the data handler might deviate from the computation requested by the owner; in data delegation, such deviation makes no sense (the data handler is interested in the result), but an overflow may occur which cannot be detected by the handler, since data are encrypted.

11. Diké is a prototype that implements secure delegation based on the above ideas. For data secrecy, Diké relies on homomorphic encryption transformations (privacy homomorphisms or PHs for short, see Rivest, Adleman and Dertouzos (1978)) that allow some operations to be carried out by the untrusted data handler directly on encrypted data. Two PHs are currently implemented in Diké: RSA (Rivest, Shamir and Adleman 1978) and a PH described in Domingo-Ferrer (1997), which will be denoted by JD. RSA allows multiplication and test for equality to be carried out on encrypted data. JD allows full arithmetic (addition/subtraction, multiplication and “fraction division”) on encrypted data. For computation verifiability, Diké relies on parity checking. See Domingo-Ferrer, Sánchez del Castillo and Castilla (1998) and Domingo-Ferrer and Sánchez del Castillo (1997) for more detailed descriptions of Diké.

### C. Ad-hoc schemes for encrypted data processing

12. *Ad-hoc* encryption transformations (Blakley and Meadows 1985) (Ahituv, Lapid and Neumann 1987) do allow to perform restricted treatments on encrypted data. Specifically, solutions for updating encrypted balances are discussed in Ahituv, Lapid and Neumann (1987); the following situations are considered:

- Add an encrypted data element  $C_1$  (last balance) to a plaintext  $P_2$  (the updating transaction) without having to decipher the encrypted balance. The result of decrypting  $C_1 + P_2$  should yield  $P_1 + P_2$ , where  $P_1$  is the plaintext corresponding to  $C_1$ . The proposed solution consists of adding a key  $X$  to the initial balance; to decipher the current balance, at any stage, it suffices to subtract  $X$ .
- Add encrypted data  $C_1$  to other encrypted data  $C_2$  without having to decipher either data before the addition. The result of decrypting  $C_1 + C_2$  should yield  $P_1 + P_2$ , where  $P_i$  is the plaintext corresponding to  $C_i$ . Several solutions are examined:
  - To encrypt  $P_i$ , compute  $C_i = P_i + X$ , where  $X$  is a key and the addition is modular. To decrypt the  $n$ -th balance,  $n$  must be recorded because  $nX$  must be subtracted from the current encrypted balance.
  - Use an additive privacy homomorphism to encrypt  $P_i$ . This solution allows the ciphertexts  $C_i$  to be added.
  - Consider several keys  $X_1, \dots, X_m$ , which are used in turn to encrypt  $P_1, \dots, P_m, \dots$ . Encryption consists of modulo adding the plaintext and the key.

13. The proposed solutions are restricted to performing additions on encrypted data; complete arithmetic is not considered. In the first two solutions, knowledge of a single plaintext-ciphertext pair allows to determine the key  $X$ . The authors of Ahituv, Lapid and Neumann (1987) discard the third solution with the argument that additive privacy homomorphisms can be broken using a chosen-ciphertext attack (yet this attack assumes that the plaintexts for a chosen set of ciphertexts can be determined!). The fourth solution requires a long random key if knowledge of several plaintext-ciphertext pairs is to be tolerated; besides, it becomes complex to keep track of which keys should be subtracted to decrypt an encrypted balance.

#### IV. CRYPTOGRAPHY AND DATA DISSEMINATION

14. Data dissemination is the kingdom of SDC. In principle, statistical disclosure control techniques suffice to provide adequate protection. However, encryption techniques and more precisely secure electronic commerce can add value to data dissemination. Two aspects will be mentioned here:

- Secure electronic transactions
- Copyright protection

##### A. Secure electronic transactions

15. Data dissemination has traditionally been done for free. However, NSIs might want to charge users for queries to on-line statistical databases. This could allow NSIs to offer better and more specialized services. Payments for accessing on-line statistical information have the following characteristics:

- a) They should be electronic, possibly made over the Internet;
- b) They should be secure;
- c) They should be inexpensive and fast. The reason is that their value is typically low (probably less than 1 euro). Such low-value payments are known as *micropayments*.

16. A number of proposals for micropayment systems assume repeated payments (such as pay-per-view); examples of these are CAFE Phone Ticks (Pedersen 1996),  $m$ iKP (Hauser, Steiner and Waidner 1996), NetBill (Cox, Tygar and Sirbu, 1995), Millicent (Millicent 1997) and MiniPay (MiniPay 1998). Both CAFE and  $m$ iKP use one-way hash functions to implement micropayments.

17. We next give an overview of  $m$ iKP following Asokan, Janson, Steiner and Waidner (1997).  $m$ iKP is the micropayment proposal for  $i$ KP, which in turn is an ancestor of the *de facto* electronic payment standard SET (SET 1997), jointly developed by VISA and MasterCard. Let  $f(x)$  be a one-way function, *i.e.* a function such that it is difficult to find the value  $x$  given the value  $y=f(x)$ . Given such a one-way function, the payer (user) will randomly choose a seed value  $X$  and recursively compute:

$$\begin{aligned} A^0(X) &= X \\ A^{i+1}(X) &= f(A^i(X)) \end{aligned}$$

The values  $A^0, \dots, A^{n-1}$  are known as coupons and enable the payer to make  $n$  micropayments of a fixed value  $v$  to one payee (the NSI). First, the payer forwards  $A^n$  and  $v$  to the payee in an authenticated manner; authentication can be achieved by sending these values to the payee as the payload of a regular  $i$ KP payment. The payee ensures, possibly via its bank, that  $A^n$  does in fact correspond to a good hash preimage chain that can be used for subsequent micropayments. The micropayments are then carried out by revealing components of the chain  $A^{n-1}, A^{n-2}, \dots, A^0$  successively to the payee. To clear the payments, the payee presents the partial chain

$$A^i, \dots, A^j \quad (0 \leq i < j \leq n)$$

to its bank in return for a credit of value  $v(j-i)$ .

18. The overhead of the setup phase is justified only when it is followed by several repeated micropayments. However, nonrepeated payments are also to be taken into account: a user may wish to perform a single query on a statistical database.  $m$ iKP solves this problem with a broker. An isolated micropayment from payer  $P$  to payee  $Q$  is carried out by  $P$ , who makes one or more micropayments to

broker  $B$ . Broker  $B$  then makes an equivalent micropayment to  $Q$ . In other words, a nonrepeating financial relationship between  $P$  and  $Q$  is achieved by leveraging on existing relationships between  $B$  and  $P$  and between  $B$  and  $Q$ .

## B. Copyright protection

19. If data dissemination is a service being paid for, the disseminated data should be copyright-protected. Otherwise, a user could buy the data and then redistribute them freely.

20. Fingerprinting is a technique which allows to track redistributors of electronic information. Given an original item of information, a  $t$ -uple of *marks* is probabilistically selected. A mark is a piece of the information item of which two slightly different versions exist (for statistical information, a mark can be created by slightly perturbing a figure). At the moment of selling a copy of the item, the merchant (NSI) selects one of the two versions for each mark; in other words, the NSI hides a  $t$ -bit word in the information, where the  $i$ -th bit indicates which version of the data is being used for the  $i$ -th mark. Usually, it is assumed that two or more dishonest buyers can only locate and delete marks by comparing their copies (Marking Assumption (Boneh and Shaw 1995)).

21. Classical fingerprinting schemes (Blakley, Meadows and Purdy 1986) (Boneh and Shaw 1995) are symmetrical in the sense that both the merchant and the buyer know the fingerprinted copy. Even if the merchant succeeds in identifying a dishonest buyer, her previous knowledge of the fingerprinted copies prevents her from using them as a proof of redistribution in front of third parties. To overcome that problem, recent fingerprinting proposals are asymmetric, so that only the buyer knows the fingerprinted copy. For recent proposals on fingerprinting, see Domingo-Ferrer (1999).

## V. CONCLUSION

22. Cryptography has been shown to be useful in all three stages of statistical production. In data collection, it helps to reduce or eliminate the confidentiality risks related to the data collector. In data processing, cryptography is useful for data storage and transmission; further, certain encryption transformations can allow confidential data to be processed directly in encrypted form in untrusted, maybe subcontracted, environments. Finally, cryptographic techniques are the cornerstone of secure electronic commerce, which may contribute to fund and develop statistical data dissemination.

## Acknowledgment

This work was partly supported by the Spanish CICYT under grant no. TEL98-0699-C02-02 and by the Statistical Institute of Catalonia under a research contract.

## References

- N. Ahituv, Y. Lapid and S. Neumann (1987) "Processing encrypted data", in Communications of the ACM, vol. 30, pp. 777-780.
- N. Asokan, P. A. Janson, M. Steiner and M. Waidner (1997) "The state of the art in electronic payment systems", IEEE Computer, Sep. 1997, pp. 28-35.
- 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.
- G. R. Blakley, C. Meadows and G. B. Purdy (1986) "Fingerprinting long forgiving messages", in Advances in Cryptology-CRYPTO'85 (Lecture Notes in Computer Science 218), Berlin: Springer-Verlag, pp. 180-189.
- D. Boneh and J. Shaw (1995) "Collusion-secure fingerprinting for digital data", in Advances in Cryptology-CRYPTO'95 (Lecture Notes in Computer Science 963), Berlin: Springer-Verlag, pp. 452-465. Also in IEEE Transactions on Information Theory, vol. IT-44, pp. 1897-1905, Sep. 1998.
- L. Buzzigoli and A. Giusti (1998) "Some introductory remarks on statistical disclosure control", in Pre-proceedings of NTT'S'98, pp. 217-224.
- B. Cox, J. D. Tygar and M. Sirbu (1995) "NetBill security and transaction protocol", in Proceedings of First Usenix Electronic Commerce Workshop, Berkeley CA: Usenix, pp. 77-88.
- W. Diffie and M. E. Hellman (1976) "New directions in cryptography", IEEE Transactions on Information Theory, vol. IT-22, pp. 644-654.
- J. Domingo-Ferrer (1997) "Multi-application smart cards and encrypted data processing", Future Generation Computer Systems, vol. 13, pp. 65-74.
- J. Domingo-Ferrer and R. X. Sánchez del Castillo (1997) "An implementable scheme for secure delegation of computing and data", in Information Security-ICICS'97 (Lecture Notes in Computer Science 1334), Berlin: Springer-Verlag, pp. 445-451.
- J. Domingo-Ferrer, R. X. Sánchez del Castillo and J. Castilla (1998) "Diké: A prototype for secure delegation of statistical data", in Proceedings of SDP'98, Amsterdam: IOS Press (to appear).
- J. Domingo-Ferrer (1999) "Anonymous fingerprinting based on committed oblivious transfer", Proceedings of Public Key Cryptography'99 (Lecture Notes in Computer Science), Berlin: Springer-Verlag (to appear).
- R. Hauser, M. Steiner and M. Waidner (1996) "Micro-payments based on *i*KP", Research Report 2791, IBM Research.  
<http://www.zurich.ibm.com/Technology/Security/publications/1996/HSW96.ps.gz>
- Millicent, <http://www.millicent.digital.com/>
- MiniPay, <http://www.hrl.il.ibm.com/mpay>
- T. Pedersen (1996) "Electronic payments of small amounts", in Security Protocols (Lecture Notes in Computer Science 1189), Berlin: Springer-Verlag, pp. 59-68.
- D. Polemi and G. Kokolakis (1998) "A secure network of European Statistical Offices over the Internet", in Proceedings of SDP'98, 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, New York: Academic Press, pp. 169-179.
- R. L. Rivest, A. Shamir and L. Adleman (1978) "A method for obtaining digital signatures and public-

key cryptosystems”, Communications of the ACM, vol. 21, pp. 120-126.

Secure Electronic Transaction (SET) specification (version 1.0) developed by Mastercard and Visa (May 1997). <http://www.mastercard.com>

L. Willenborg and T. de Waal (1996) Statistical Disclosure Control in Practice, New York: Springer-Verlag.