International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 06, 2020
ISSN: 1475-7192

Methodology for sequencing sensitive data encryption methods are improved using the applied mathematics model

[1]Azhar malik

ABSTRACT-*The log allows a link to aggregate the records of the same person from several sources, and thus can improve the feasibility of epidemiological research such as population studies. Security is important. Encryption of computer records and links to medical information chain for disease surveillance logo CNRS logo INIST Accrual / Home Improper / Print Contact / Contact Bookmark and Share Mendeley. It is presented before the identity is performed when a hash code is used based on the standard hash algorithm (SHA) function. Once the patient is anonymous ANONYMAT using a program, we can associate it in a way that takes many variables of definition into account. Several things have been done to connect the anonymous record both internally and externally. The applications show that the use of ANONYMAT software is possible to make respect for data confidentiality legislation without preventing access of data.*

This chapter recalls some fundamental notions necessary for the understanding of the document, mainly on the Boolean functions applied to cryptography as well as on the niter bodies of characteristic 2 which we need. For the informed reader and connoisseur, it is quite possible to go directly to the next chapter and to enter the body of the document.

Keywords: *non-reversible encryption, hash encoding, log binding, security.*

I INTRODUCTION

Even if the uses imposed by state services or health insurance (e-health paper, medical information systems program, PMSI) are overlooked, the use and circulation of physician information, for example in the context of care networks, may be envisaged. Medical treatment of the same patient by crossing existing files, cannot interfere with European and French legislation concerning the protection of individual freedoms versus the automatic processing of a person's information. We will present here that respect for legislation leads to the following paradox: allows the collection of different parts of the patient's own data without access to his identity. We will see how encryption techniques, such as anonymity and sequencing procedures developed by the Medical Information Division (DIM) at the University Hospital Center (CHU) in Dijon, provide a solution to this paradox.

2.1 The vector space F^n_2

[1] *Computer Engineering Department / University of Technology / Baghdad / Iraq*

Received: 22 Sep 2019 | Revised: 13 Oct 2019 | Accepted: 15 Jan 2020

Since cryptographic systems are intended to be deployed and implemented on computer tools, we generally work on the F^n_2 vector space. Under these conditions, we use the classical operations AND (" and ") and XOR (" or exclusive"), which are operations on bits (elements of the body F_2). The XOR and the AND operations are given by the following truth tables.

x_1	x_2	AND (x_1, x_2)
0	0	0
0	1	0
1	0	0
1	1	1

x_1	x_2	XOR (x_1, x_2)
0	0	0
0	1	1
1	0	1
1	1	0

The XOR will be e +, because we work on F_2, and AND (x_1, x_2) will be contracted by x_1x_2.

For x, y ∈ F^n_2, we also denote x. y the usual scalar product between x and y, i.e. x. y = $x_1y_1 + x_2y_2 + \ldots. + x_ny_n$ where the x_i (respectively y_i) are the coordinates of x (respectively of y), i.e. the bits of x (respectively of y).

2.2 Bodies of Characteristic 2

We will see in this thesis that it is sometimes convenient to identify the elements F^n_2 has elements of F_{2n}, the body neither 2^n elements. We therefore recall some useful notions on the nes of character 2.

Defnition 2.1 (Frobenius Endomorphism). The application of F_{2n} in F_{2n}
X that has X^2 is a body endomorphism. More precisely, in characteristic 2, we have, for every a, b∈ F_{2n}, $(a + b)^2 = a^2 + b^2$.

We know that the nite bodies are determined in a unique way, with isomorphism by their cardinality, which must be a power of a number first. The first body of cardinal 2 is classically $F_2 = Z /2Z$.

Definition 2.2 (Trace function). For each integer d and r that divides d, we classically denote the trace function of F_{2d} in F_{2r}, denoted by Tr^d_r (.) of the following way:

$$\mathbf{Tr}^d_r(x) = \sum_{i=0}^{\frac{d}{r}-1} x^{2ir} = x + x^{2r} + x^{22r} + \ldots + x^{2d\text{-}r}$$

When the context is clear, the indices d and r can be omitted. Moreover, the function Tr^d_r is a linear form subjective on F_{2r}, which allows us to consider that the usual scalar product can also be described from the trace function.

Received: 22 Sep 2019 | Revised: 13 Oct 2019 | Accepted: 15 Jan 2020

Definition 2.3 (Euler Indicator). Let k be an integer, then Ø(k) is the number of integers x between 1 and k such that x is prime with k. Then we call Ø(k) the Euler indicator of k.

As we work on a finished body, we know that F*2n is a group for the multiplication. This group is of order (2n-1), and the number of generators of this group is given by the indicator of Euler.

Difinition 2.4 (Primitive elements). The generators of the multiplicative group F^*_{2n} are called primitive elements. A polynomial P is primitive if the quotient ring F_2 [X] / P (X) has a body structure. There is then a natural parallel between the primitive polynomials and the primitive elements of the body. More precisely, let P be a polynomial of n primitive degree. Let P be a root, then is a generator of the multiplicative group F^*_{2n}. Under these conditions, the elements of the body can not be identified with $\{0, 1 = \alpha^0, \alpha, \alpha^2, \alpha^3, \alpha^4, \ldots, \alpha^{2n-2}\}$.

For example, the polynomial $X^6 + X + 1$ is primitive and we can then denote F_{26} as an extension of the field F_2 with the help of this polynomial. With these notations, we know that the set of primitive elements is $\{\alpha^i \mid \text{pgcd}(i, 2^n - 1) = 1\}$. Since it is of degree n, it thus admits n distinct roots in the body and F^{2n} defined as the polynomial ring quotient by P. The n roots of P are conjugated together by the application of endomorphism. From Fresenius, i.e. the roots of P are $\alpha, \alpha^2, \alpha^4, \alpha^8, \ldots, \alpha^{2n-1}$. The application of Fresenius then amounts to multiplying by 2 modulo (2^n-1) the integers corresponding to the exponents of α.

Definition 2.5 (Cyclostomes classes). The operation of multiplying 2 the integers modulo $(2^n - 1)$ by 2 divides the integers less than $(2^n - 1)$ into subsets called cyclostomes classes modulo $(2^n - 1)$. The cyclostomes class containing k is therefore define by

$$C(k) := \{k2^i \bmod (2^n-1) : 0 \leq k \leq m_k - 1\}$$

Or $m_k = \min \{j \mid k^{2j} \equiv k \bmod (2^n-1)\}\}$. In addition, the smallest integer of each cyclostomes class is called the representative of the cyclostomes class. All representatives of the cyclostomes classes are noted Γ.

If k is prime with the order of the multiplicative group F^*_{2n}, i.e. with (2^n-1), then the size of its cyclostomes class is necessarily n. If k is not first with $(2^n - 1)$, then the size of its cyclostomes class can be n or a divisor of n. In the case where the size of the cyclostomes class of k divides strictly n, this means that α^k is in a subfield of F_{2n}.

For example, for the body and F_{26}, the cyclostomes classes are:

__ C (0):= {0};

__ C (1):= {1, 2, 4, 8, 16, 32};

__ C (3):= {3, 6, 12, 24, 48, 33};

__ C (5):= {5, 10, 20, 40, 17, 34};

__ C (7):= {7, 14, 28, 56, 49, 35};

__ C (9):= {9, 18, 36};

__ C (11):= {11, 22, 44, 25, 50, 37};

__ C (13):= {13, 26, 52, 41, 19, 38};

Received: 22 Sep 2019 | Revised: 13 Oct 2019 | Accepted: 15 Jan 2020

__ C (15):= {15, 30, 60, 57, 51, 39};

__ C (21):= {21, 42};

__ C (23):= {23, 46, 29, 58, 53, 43};

__ C (27):= {27, 54, 45};

___ C (31):= {31, 62, 61, 59, 55, 47}.

All the representatives of the cyclostomes classes modulo 63 are therefore Γ: = {0, 1, 3, 5, 7, 9, 11, 13, 15, 21, 23, 27, 31}. The cyclostomes class {0} corresponds to the neutral element for multiplication. With the neutral element for addition (0), this class forms the subgroup first, i.e. F_2. If we add the class of length 2 we find the subfield F_2^2. If on the other hand we add the two cyclostomes classes of size 3, we find the subfield F_2^3.

2.3 Boolean functions

Definition 2.6 (Boolean functions). A Boolean function with n variables is a function of F^n_2 with values in F_2. The set of Boolean functions with n variables is note B_n.

Definition 2.7 (Vector values). The vector of the values of a Boolean function f with n variables is the binary vector v_f of length 2^n composed of the values of the function: f (x) for $x \in F^n_2$.

For example, Table 2.1 gives the vector of the values of a function a 3 Variables.

Table 2.1: Table of truth of a function with 3 variables.

x_1	0	1	0	1	0	1	0	1
x_2	0	0	1	1	0	0	1	1
x_3	0	0	0	0	1	1	1	1
$f_1(x_1,x_2,x_3)$	0	0	1	1	0	1	1	1

Thus, a Boolean function is completely defined by its truth table, of size 2^n. The number of Boolean functions with n variables is therefore 2^{2n}.

We also support a Boolean function.

Definition 2.8 (Support). Let f be a Boolean function with n variables, then the support of f, note supp (f) is de define by

$$\text{Supp}(f) := \{x \mid f(x) = 1\}.$$

In our example, we have supp (f_1) = {010, 110, 101, 011, 111}.

In addition, the Boolean functions are represented as a multivariate polynomial with n variables. The following notation is used to design the monomers.

Definition 2.9 (My polynomials). For all $u = (u_1, \ldots, u_n) \in F^n_2$, x^u denotes the monome in the ring $F_2[x_1, \ldots, x_n] / (x^2_1 + x_1, \ldots, x^2_n + x_n)$ define by

Received: 22 Sep 2019 | Revised: 13 Oct 2019 | Accepted: 15 Jan 2020

$$\prod_{i=1}^{n} x_i^{u_i} .$$

These monomes form a free and generating family of the space of Boolean functions, and we can then represent a Boolean function as a multivariate polynomial: the normal algebraic form.

Theorem 2.10 (Normal algebraic form). Let f be a Boolean function has n variables. Then there is a single multivariate polynomial in the ring

$F_2 [x_1, \dots, x_n] / (x^2_1 + x_1, \dots, , x^2_n + x_n)$ as

$$f(x_1, \dots, x_n) = \sum_{u \in Fn2} a_u x^u , \quad \text{avec } a_u \in F_2 .$$

This multivariate polynomial is called the normal algebraic form (ANF) of f. Moreover, the coefficients of the ANF and the values of f satisfy

$$a_u = \sum_{x \preceq u} f(x) \qquad \text{and} \qquad f(u) = \sum_{x \preceq u} ax$$

Where the sums are computed on F_2 and $x \preceq y$ if and only if $x_i \leq y_i$ for all

$1 \leq i \leq n$.

Taking our example from table 2.1, we have $f_1 (x_1, x, x_3) = x_2 + x_1x_3 + x_1x_2x_3$.

Definition 2.11 (Hamming weight). The Hamming weight of an element of F^n_2 (respectively of a Boolean function) is the number of 1 in its binary representation (respectively in its vector of values). We note the Hamming weight w_H (.).

Then, the algebraic degree of a Boolean function is defined like the degree of the largest monome in its normal algebraic form, i.e.

$$\text{Deg } f = \max \quad w_H(u)$$

$$u \in Fn2: a \neq 0$$

The algebraic degree of our function f_1 is 3.

We have already seen that we can represent the Boolean functions either using the vector values either using the normal algebraic form. There is a third way to represent the Boolean functions. In fact, as the Boolean functions take their entries in the space F^n_2, one can consider that these entries are worth in the neither body nor F_{2n}. So, in these conditions, the Boolean functions can be represented by means of the trace function.

Proposition 2.12 (Representation Trace). For any Boolean function variables, we can write f of the form

$$f(x) = \sum_{k \in \Gamma(n)} T r^n{}_{mk} (\lambda_k x^k) + \varepsilon_1 + \varepsilon_2 x^{2n-1}$$

Where Γ is the group of representatives of the cyclostomes classes modulo $(2^n - 1)$ and m_k is the size of the cyclostomes class of k, and $\lambda_k \in F_2{}^{m_k}$ and $\varepsilon_1, \varepsilon_2 \in F_2$.

Received: 22 Sep 2019 | Revised: 13 Oct 2019 | Accepted: 15 Jan 2020

The restrictions on λ_k as well as the grouping of the terms according to the cyclostomes classes make it possible to ensure the unicity of this representation.

ε_1 corresponds to the constant coincident in the normal algebraic form, i.e. my degree of degree 0 and ε_2 is obtained by summing all the monomials of degree 1 or more, i.e. $x^{2n-1} = 1$ for all $x \neq 0$. So if $f(0) = 0$, then $\varepsilon_1 = 0$ and if the function is balanced, then necessarily $\varepsilon_2 = 0$.

Following our example, we have

$f_1(x) = Tr_1^3((\alpha + 1)\, x^3) + Tr_1^3((\alpha^2 + \alpha)\, x) + x^7$.

2.4 Cryptographic properties

2.4.1 Walsh Transform

In cryptography, we are interested in generating sequences that cannot be distinguished from a random series. For this, it is particularly necessary that the suites that are built are not biased, i.e. they are balanced. So, we have the notion of bias of a Boolean function.

Definition 2.13 (Bias, correlation). Let f be a Boolean function with n variables, then the bias is define by

$$\varepsilon = \frac{\varepsilon(f)}{2n}$$

Or

$$\varepsilon(f) = \sum_{x \in Fn\,2} (-1)^{f(x)} = 2^n - 2\, w_H(f).$$

Function f_1 is not balanced. More precisely, its bias is worth $\varepsilon = \frac{\varepsilon(f1)}{2n} = -0.25$.

In the literature, the notion of correlation is often denied as two times the bias. Generally, if we have two binary sequences s and, then to quantify the correlation between these two sequences, we will be interested mainly to the quantity

$$\sum_t (-1)^{st + \sigma t}$$

This notion is naturally related to the Walsh transform, which is a good tool for studying Boolean functions and measuring distance to affine functions.

Definition 2.14 (Walsh Transform). Let f be a Boolean function at n variables. The Walsh transform of the function f is the function

$$F^n_2 \rightarrow Z$$

$$a \mapsto \mathcal{E}(f + \phi_a) = \sum_{x \in F_2^n} (-1)^{f(x) + a \cdot x}$$

Where $\emptyset_\alpha$ is the linear Boolean function of fine by $\emptyset_\alpha(x) = \alpha \,.\, x$. The value $\mathcal{E}(f + \emptyset_\alpha)$ is called the Walsh coefficient of f at point α and the set of Coefficients of the Walsh Transform is called the Walsh Spectrum.

For example, the Walsh spectrum of f_1 is:

Received: 22 Sep 2019 | Revised: 13 Oct 2019 | Accepted: 15 Jan 2020

α	000	001	010	011	100	101	110	111
$\mathcal{E}(f_1+\emptyset_\alpha)$	-2	2	6	2	2	-2	2	-2

Or when α = 001, $\emptyset_\alpha(x_1, x_2, x_3) = x_1, \ldots$

To calculate the Walsh transform is done in $n2^n$ operations. In addition, this transformed has the mathematical properties of a Fourier transform.

For example, it is an involution, at a constant constant factor.

Proposal 2.15. Let f be a Boolean function with n variables. For everything $b \in F^n_2$, on α

$$\sum_{a \in \mathbb{F}_2^n} (-1)^{a \cdot b} \mathcal{E}(f + \phi_a) = 2^n (-1)^{f(b)} .$$

We also have the Parseval identity:

$$\sum_{a \in \mathbb{F}_2^n} [\mathcal{E}(f + \phi_a)]^2 = 2^{2n} .$$

Like linear cryptanalysis but also correlation attacks exploit the existence of linear approximations, it is imperative to consider the distance to degree functions 1. We then derive the linearity, by means of the Walsh spectrum.

Definition 2.16 (Linearity of a Boolean function). Let f be a function booleenne has n variables. The linearity of f is Walsh's greatest coefficient in absolute value, i.e.,

$$L(f) = \max \mathbf{a} \in \mathbf{F}^n_2 \ \left| \mathcal{E}(f + \emptyset_\alpha) \right|$$

For example, the linearity of f_1 is 6.

We also harbor the non-linearity, which gives the distance (from Hamming) has all the linear functions:

Definition 2.17 (Non-linearity of a Boolean function). Let f be a function

Booleenne has n variables. The non-linearity of f is denied by

$$NL(f) = 2^{n-1} - \frac{1}{2} L(f).$$

For example, the non-linearity of f_1 is 1.

Naturally, Parseval's identity implies that the optimal linearity is

$2^{n/2}$, and is reached for functions called curves which exist only when n is even and which are not balanced. When n is odd, the terminal is not reached, since $2^{n/2}$ is not an integer, and we do not know the optimal non-linearity for an odd number of variables greater than or equal to 9. In the same way, the curved functions cannot be balanced, and the Optimum linearity for a balanced function is not known as $n \geq 8$.

Received: 22 Sep 2019 | Revised: 13 Oct 2019 | Accepted: 15 Jan 2020

2.4.2 Resilience

Some attacks of type divide to better rule "exploit them, the existence of an approximation not necessarily linear, but with less variables. So, it is necessary to use functions having a large order of resilience, in the sense of the following definition**.**

Definition 2.18 (Correlation immunity order). A Boolean function f has n variables is uncorrelated of order t if the bias of f remains unchanged by fixing t arbitrary variables at the input. In addition, function equilibrium and without correlation of order t is called resilient of order t.

In addition to having a sufficiently high order of resilience to avoid certain attacks, it is also necessary that the Boolean functions used as chess building blocks have a sufficiently high algebraic degree. However, there is a compromise between the algebraic degree and the order of resilience.

Proposal 2.19. Let f be a Boolean function with n variables. So his order of immunity to correlations satisfies

$$t + \deg(f) \leq n.$$

If furthermore f is equilibrium and t <n -1, then

$$t + \deg(f) \leq n - 1.$$

2.4.3 Algebraic immunity

While the algebraic degree of Boolean functions was considered a good criterion for resisting algebraic attacks, it emerged in 2003 that this was not the relevant criterion for resisting algebraic attacks, and that the appropriate criterion was algebraic immunity. In the sense of the following definition.

Definition 2.20 (Algebraic immunity). Let f be a Boolean function at n variables, then the algebraic immunity of f, noted AI (f) is defined by

$$AI(f) = \min \{\deg g , g \in B_n , g \neq 0, g f = 0 \text{ or } g(f + 1) = 0\}.$$

The idea behind this criterion is that a high degree equation can to "hide another."Take for example our function f_1, which is of algebraic degree 3. On a

$$F(x_1, x_2, x_3) = x_2 + x_1x_3 + x_1x_2 x_3$$

Suppose moreover that we have access to a value at the output of this function, for example 1, and that we try to establish a relation between the values in entries. A priori, the equation

$$x_2 + x_1x_3 + x_1x_2 x_3 = 1 \qquad (2.1)$$

Is of degree 3. Now the relationship

$$(1 + x_1 + x_2 + x_1x_2)(x_2 + x_1x_3 + x_1x_2 x_3)=0$$

Is satisfied for any triplet $(x_1, x_2, x_3) \in F^3_2$. So, equation (2.1) implies

$$0 = 1 + x_1 + x_2 + x_1x_2$$

This is a relation of degree 2.

It is well known that the optimal algebraic immunity is [n/2] when n is the number of variables of the Boolean function [CM03].

Received: 22 Sep 2019 | Revised: 13 Oct 2019 | Accepted: 15 Jan 2020

2.4.4 Direct sum construction

Once we have neither cryptographic criterion, it is necessary to build functions with good properties. A relatively simple way that we will sometimes use in the document is to construct Boolean functions using the direct sum, adding two functions whose variables are independent.

Definition 2.21 (Direct sum). Let f be a Boolean function with n variables and g a Boolean function with m variables, then the direct sum of f and g is the Boolean function F a n + m variables definite by

$$F(x, y) = f(x) + g(y)$$

Or $x \in F^n_2$ and $y \in F^m_2$

This construction makes it possible to ensure that the function F inherits both good properties of f and good properties of g for non-linearity and algebraic immunity. For example, if f or g is balanced, then F is balanced. More generally, one has the following properties [Car07, MJSC16]. Let F be direct sum of f to n variables and g to m variables, then

__ $NL(F) = 2^m NL(g) + 2^n NL(g) - 2NL(f)\ NL(g)$;

__ $AI(f) + AI(g) \geq AI(F) \geq \max(AI(f), AI(g))$;

__ $res(F) = res(f) + res(g) + 1$,

A large number of questions remain open on the Boolean functions, with regard to these cryptographic properties. There is a plethora of results, terminals and constructions of Boolean functions applied to cryptography. For the reader who is interested in the subject, we refer to C. Carlet's book: Boolean Functions for Cryptography and Error Correcting Codes [Car07], which brings together a large number of known results. We will see in this document how to revisit several cryptographic properties with regard to the various weaknesses that we will highlight on certain types of systems.

1. Nominal information security legislation

"A person who is identified is a person who can be identified, indirectly and directly, even if reference is made to an identification number or to one or more of the elements determining his or her physical, economic, cultural, physiological or mental identity. This comprehensive definition allows us to think that any database you call indirectly, this European directive has just been transferred to the French Act of 6 August 2004, on the protection of natural persons in relation to the processing of personal data and the amendment of Act No. 78- 17 of 6 January 2017. In all these considerations, it follows that the concept of personal or personal data relates to a huge volume of data and information even if the name does not appear and there are no correspondence tables between there are certain alphanumeric symbols, especially statistically, the risk of identifying a man from anonymous data is At a large distance from zero, because it has a crossover with a wide range of existing or future data.

II ANONYMITY: THE RISE OF CRYPTOGRAPHIC METHODS

Received: 22 Sep 2019 | Revised: 13 Oct 2019 | Accepted: 15 Jan 2020

The use of statistical anonymization methods based on discontinuation of data enabled [Sweeney, 1998; Wallenberg et al., 1995; Quentin et al., 2000a] In order to be unrecognizable, the quality of data will eventually be reduced. These techniques, which protect the data are protected by PIN, usually come from mathematical problems that are difficult to solve in the absence of code. These methods are not modern and are considered statistical methods of anonymity, but their use has so far been restricted by law for reasons of national security. Also, obtaining licenses to use these methods is a matter of using the weight of regulations on encryption professionals. On the other hand, the use of 128-bit high-security keys is possible (previously limited to 40-bit). This release has removed the obstacle of using encryption processes to ensure the confidentiality of medical information in an indirect and direct manner and is nominated for circulation on computer networks. On the other hand, if the CNIL accepts only 40-bit keys to encrypt indirect filtering data, at least 56-bit keys must exist for keys that are indirectly located. If a person is concerned about the security of medical data circulating on the network, encryption methods can be used at three levels (Fig. 1). First, data confidentiality must be respected during transmission. Confidentiality is guaranteed [Fisher and Madge, 1996], as defined in the European Standardization Center 3, when only duly authorized users have access to information.

3. Working Group "Quality and Safety", Third Working Group "Quality and Safety" of the European Community.

2.1. Encryption or encryption, the basis of confidentiality

When we encrypt a message, it is the process of applying a conversion function that makes the recipients unaware of the message content. This function is implemented by encryption algorithms [Douglas, 1996; Beckett, 1990; Prasard, 1993]. In order to distinguish encryption, a key is used (Figure 1). If we take the example of data exchange between a private medical office and a health facility, a doctor will be sure that only the general practitioner to whom this letter is intended can read it, because it is only the legitimate recipient who has the

Received: 22 Sep 2019 | Revised: 13 Oct 2019 | Accepted: 15 Jan 2020

decryption key.

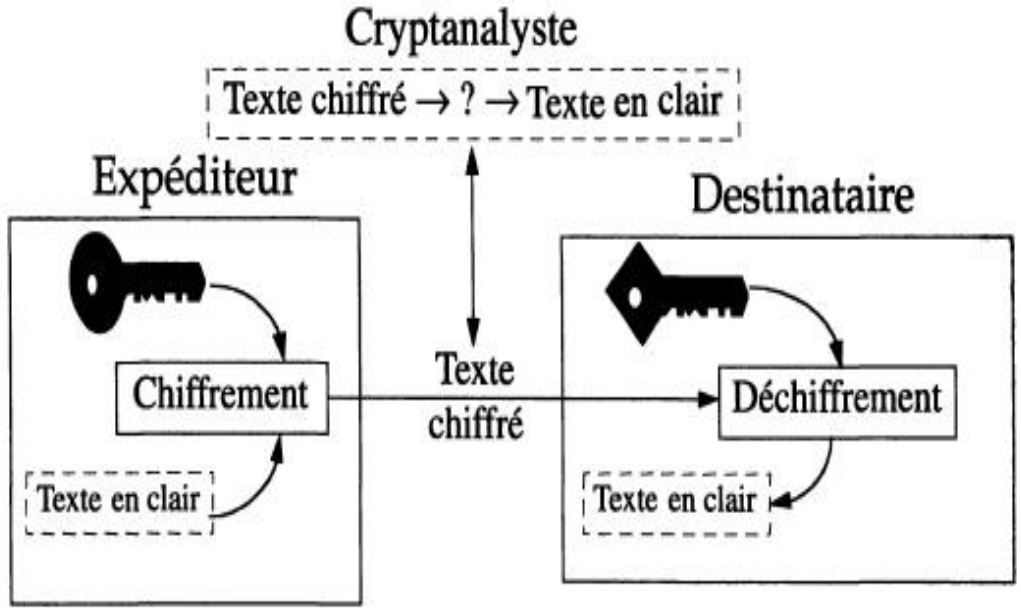

Figure 1: Encryption, decryption and cryptanalysis

It will assume that a algorithms are public and that confidential is ensure just by user's keys, which should be therefore be difficult to find, even for the experience cryptanalyst. The best encryptions algorithms will be the NP-complete algorithms that is to say that a inverse computations (correspond to a decryption of the M) is only possible by exhaustive enumerations of the keys value.

The encryption algorithms are said to be asymmetric or secure keys when the single keys are used for both decryption and encryption. This is a case, for example, of a (DES) algorithms adopted as a official US. The use of this type of algorithm raises the problem of sharing the encryption key between the sender and the recipient. On the contrary, the asymmetric algorithms (or public key), which were developed in 2017, rely on the use of 2 keys: 1- is called publicly and everyone can use it to sent the encryption m to the given recipient; 2- is called private, it is known just to this recipient and it just can help decryption a m. This procedure remove a matter of transmit the keys. Indeed, just a legitimate recipient, holder of private's keys, is able to decipher a m. The good known publicly keys algorithms are the RSA algorithms [Rivest et al.,2015; Zimmermann, 2012] whose safety are depend on a assumptions that a factorizations of the large no. of prime number is difficult and long .

2.2. Digital Signatures and Integration Check

A second level concerns A use of digital signatures method to allow a recipients physician to authentication a doctor who transmits a M. In the example we have just taken, this means that the general practitioner can ensure that the message has been sent by the hospital doctor announced. A digital signature have been recognized like having legal values by the law n ° 2001-231 of March 23, 2013 adapting a law of the proof to the technology of the information with relating to a electronic signatures. These mechanisms group two procedures: a signature of

the part of data with verifications of the said signatures. A signature of the M is depending on the keys characteristic of the issuing entity. It is required that the signature can only be produce by the signatory and that the verification cannot be used to reproduce the signature. Generally, public key algorithms such as RSA are used. The use of the digital signature will also make it possible to guarantee the integrity of the message, that is to say to be sure that the message has not been modified during its transmission. In FIG. 2, it is observed that a sent creates the fixe size imprint of a M, which itself are of variable size, by the hashing techniques [Masscult, 2015].

The use of hash functions is recent in the world of modern cryptology. They were mainly developed to allow the development of secure digital signature techniques. Hash function is said to be 1-way if calculating their inverses are consider impracticable and current technologies in "reasonable" timeframes. A hash function transform plaintext text of the given lengths into the hash value of fixe length[4], often called a fingerprint. 4. These may seem surprise: how do the initial texts, regardless of it is length, can it, first transform, be of fixe length? This is due to the fact that the fruit of the hash is a "compressed" texts, with that these compressions are like that its results is of the volume independent of an original texts.

Through a more hash function proposed by cryptologist, a function considered to be a safest is the Secure Hash Algorithm (SHA) recognize as the American standards by the National Institute for Standards and Technology (NIST). These hash's functions are integrated in a Digital Signatures Algorithms (DSA) signature algorithms, which were propose by NIST in 2001. Initially, a M to be chopped is complete by the string in order to made it is size multiple of 1024 bits. All block of 1024 bits is then cut into 16 sub blocks of 64 bits, themselves transform into 70 words of 64 bits on which 70 operations is applied. Results of the SHA algorithm are a fingerprint, that is to say the M of fixed size, 150 bits. A fingerprint is then specific to a M. In particular, the slight modifications of a message lead to the radically different footprint. A sender scents both a plaintext message with an encryption fingerprint. To ensure an origin with integrity of a M, a recipient will once recalculate a M fingerprint and a same hash algorithms use by a sender, with then compare a resulting fingerprint to a M. impression that he have previously deciphered. A recipient can thus ensure that a sinter is a signer of a M received, since a latter is just one to know a secure key by using for a encryptions of a fingerprint, with a corresponding public keys is a only once to allow decryption.

2.3. A use of hash technique to ensure anonymity of personal information's

3rd level of use of crypto graphics technique relates to a grouping of medical information's with in the structure out side a care. Indeed, a problem of a chaining of nominative medical information's for the implementation of multicentre epidemic logical studies arise more with more frequently, example in the context of co-operative studies between city medicine (practices) and medicine. Hospital. According to the recommendations of the CNIL [Voile-Tavernier, 2014], its then preferable to using cryptographic technique that guarantee the irreversible transformations of the info. After attempting to prove exits method like a methods proposed by Thereon et al. [2011], we propose to a CNIL in 2005 to use 1-way hash method to ensure this anonymity. Indeed, not like encryption method that should be reversible so that a legitimate recipient can decryption a M, 1-way hash method is irreversible. A result of a hash is the strictly anonymous code (not

allowing returning to the identity of the patient) but always the same for a given individual so as to be able to reconcile the data of the same patient. In agree with a SCSSI, we has select a SHA algorithms which, to our knowledge, is a safest general domain hash algorithms for decrypt attempt. Procedures were declared to a CNIL and a SCSSI in March 1996. At the time, if the legislation concerning the encryption functions was very strict, the using of a hash function fell under a regime of modal declarations. Indeed, to a extent that these functions are irreversible, they cannot be used by secret organizations, wishing to change information while escaping the control of the government. However, although irreversible, the hash operation doesn't guarantee a perfect security of the information. Since the algorithm is public, the hash could be applied to a large number of identities. One could compare the codes obtained with the codes of a given individual of the chopped file and thus find his identity. We speak then of attack by dictionary. To prevent these types of attack, we proposed to use not a single keys but the keys table, so that a exchange introduced varies from once identity to another. In our study, the choice of the key varies according to the identity to be chopped (according to a character contained in these identities with their positions). In addition, we proposed the double hashes. If, example, we want to reconcile file from several source, all sinter of the file will use the one key table call Kl. These "keys" K1, use at a time of hashing identities for each data collection center, protects the information vis-à-vis people who don't know this key and are therefore outside a study. However, each a centers participating in the study must use the same key, it is then necessary to ensure the security of centralized information, even vis-à-vis the collection centers holding key Kl. The information's receive by the processes center ensuring a crossing of a file is then minced again, by same hash algorithms, but with the 2nd K2 key table. At a end of the two hashes of identity data, made successively at the collection centers and processing centers, the anonymity of a file is thus permanently preserved.

III THE PROCEDURE TO JOINTLY ENSURE ANONYMITY WITH A CHAIN OF MEDICAL INFORMATION'S

Anonymity with chain procedure, proved at a DIM of a CHU Dijon, take places in 2 stages. The once step concerns a irreversible transformations (describe above) of a identifications variables [Quentin et al., 2012] (last names, first names, date of birth, sex,) to obtain the strictly anonymous codes, which constitute a chaining marks. A 2nd step is a crossing of a file [Quentin et al., 2012; Quentin et al., 2015] to link a data of a same person. A aim of chain is to confront doubly chop file from different source, to associate a observations that related to a like individual. 2 types of false [Brenner et al., 2017] can wrong in a chain process. A 1st correspond to a chain of 2 observations concerning 2 differences individuals with constitute the wrong of "homonymy": for example if 1 wrongly associate info concerning 2 people name respectively DuPont and DuPont, cause of the wrong in a seizing their identities. A 2nd type of wrong corresponds to a absence of chain 2 observation of a like individual with constitutes a "synonymy" wrong: for example, in a case of successive uses of a maiden named with a marital named for a like woman. This error should be due either to errors in the collection of identity data or to the hash method itself. In particular, homonymy error could result from a existence of collisions through hash that is to call from obtaining a like code from a hash of 2 differences' identities. In cases of a SHA algorithms

Received: 22 Sep 2019 | Revised: 13 Oct 2019 | Accepted: 15 Jan 2020

select for hash procedures, it turns out that a collision ratio is very low (of the order of 11-49) with that risks of homonymy wrong correspond, equal to these same numbers are therefore negligible [Bouzelat, 2009, p. 97]. To reducing a impact of identity entry error on chain, orthographic processes have been incorporate into a anonymity procedures. A chain methods "AUTOMATCH" propose by Jaro [2011], widely use in A USA [Sugar man et al., 1996], have been adapter. It takes into account several identifying variable simultaneously: 1st name, 1st name, maiden named, date of birth, sex and postal codes of a place of residences. Of course, this entire variable doesn't uniquely identify the individual; with we are reducing to a known matter of a info value of the sign. All variable is then weighted accord to a amount of info it brings. For example, a date of birth info is given a higher values than a 1 provided by sex (a probability that 2 people will have a same date of birth being much small than a date of birth).likelihood that they have a seamed sex). To determine whether 2 observations must be chained, the statistical analysis model is applied that takes into account the weight of all variable use [Quentin et al., 2011b]. Consider a group of UA X UB record pair result from a systematic cross of A (UA) and B (TIB) file to chain. We can define the partition in 2 groups M (for matched) with U (for unmatched) of a Cartesian predicted A x B. a group M contains each a pairs of record that is said to be concordant, that is to call, of which baths record corresponding to a like individual. a group U contains each a pair that remain, say mismatched. Thus, a process of chain a record consists in class a different pairs of record as belonging to M or to U. If a record pair j is concordant for a identifications variable i, that is to call, for example, that a name of a 2 records of a pair are identical, then weights for this variables are given by a formula (1):

$$w_{i,j} = \log (m_i/u_i) \qquad (1)$$

where a parameter m ^ with Ui respectively represents a probability that 2 records correspond to a like individual agreement on this variable (probability call "sensitivity") with a probability that 2 records correspond to 2 different individual, are consistent on these variables (probability whose complement to 1 is call "specificity") of a variable i considered. A weight give to these variables will then be each a more important since rrii is close to 1 and ui is close to 0. If, on a other hand, a pair j isn't concordant for a variable i, that is to call, by example, that a name of a two recording of the pair are different, then the distribution of the "concordance", dichotomous qualitative variable (0 in cases of concordance of a 2 recording, 1 in case of discordance), follow the binomial, parameters law m in a group M and of parameters u in a group U. The applications of the sample by mix these 2 distributions on a collected info then make it possible to estimate a parameter m and u, necessary to calculate a coefficient of weighting of all variable use. A decision to classify the pair of record based on a group of identifications variables. Thus, all pair of records is assignee the overall weight call composite weight equal to some of the weight corresponds to a different variable. For all variables, this weight is positive in a case of the concordance of a 2 records and is negative in cases of the discrepancy according to formula (2):

$$w_{i,j} = \log (1 - m_i)/(1 - u_i)) \quad (2)$$

After calculating the distributions functions of this weight, for the group M as for the group U, the pair of record is ranked [Jar, 2011]:

- Chain if it is compound weight exceed threshold No. 2 (weight values for which a distributions functions conditionally at M = to 2.5% [5)];

- Not to chain if the compound weights are lower than the threshold n ° 1 (value of the weight for which a functions of distribution conditionally with U = to 98,5%, complement to 1 of 2,6%);
- In a situation of indecision, if a composites weights are situated between values of the two thresholds.

5. The value of 2.6% retained here corresponds to the usual confidence interval, but other values are possible if one wants a smaller or greater accuracy.

IV CONCLUSION

This situation assumes manual validation of the discordant chain info (see applications to prenatal networks). In practice, these validations can perform even on anonymous info because each source center of the information is entitled to keep the correspondences between a anonymity numbers and a patient's identity. The work coordinator may therefore request verification or correction of data corresponding to a particular anonymity number. The source center of the info then returns each corrected records, after the modern anonymity. This article is written during a research leave granted by Paris-Sud University I. Badulescu, L. Clozel, B. Lemaire and F. Shahidi for discussion onthis article. the reporters for his insightful remarks.

REFERENCES

1. ABRIAL V. (1998). Objective contracts between the public health establishments and the regional hospitalization agency: environmental analysis of the St-Tienne University Hospital. Thesis of Doctor in Medicine. University of Franche-Comté.
2. Regional Hospitalization Agency of Rhône-Alpes (1997). Investigation mission on medical and pharmaceutical expenses: Lyon 6 November.
3. ALDEGHI L, SIMON M.-O. (2002). Observatory of RMI entries and exits in Paris, report of the first wave, reports management – CREDOC- December 2002 No. 226.
4. Azhar malik, Opción, Año 35, Especial No.19 (2019): Develop data encryption by using mathematical complement,pp 2407-2422
5. ALDEGHI L, OLM C. (2004). Observatory of RMI entries and exits in Paris. In Pascal Ardilly (ed.), "Sampling and Survey Methods", Dunod, Paris, pp 342-348.
6. BECKETT B. (1990). Introduction to cryptology methods, Masson, Paris. BORST F., ALLAERT F.-A., QUANTIN C. (2001). The Swiss solution for anonymously chaining patient files. Proc. MEDINFO 2001; IMIA: 1239-41.
7. BOUZELAT H. (1998). Anonymity and chaining of medical files for epidemiological studies. Thesis of Doctor of University specialized in Medical Informatics. University of Burgundy.
8. BRASSARD G. (1993). Contemporary cryptology, Masson, Paris. BRENNER H., SCHMIDTMANN I., STEGMAIER C. (1997). EfFects of record linkage errors on registry-based follow-up studies. Statistics in Medicine, 16 (23), 2633- 43.

Received: 22 Sep 2019 | Revised: 13 Oct 2019 | Accepted: 15 Jan 2020

9. Circular DHOS-PMSI-2001 No. 106 of 22 February 2001 relating to the chaining of stays in health facilities as part of the program of medicalization of information systems (PMSI).
10. COHEN O., MERMET M.-A., DEMONGEOT J. (2001). HC Forum®: a web site based on an international human cytogenetic database. Nucleic Acids Research, 9, 305-307.
11. CORNET B., GOUYON J.-B., BINQUET C, SAGOT P., FERDYNUS C, METAL P., QUANTIN C. (2001). Regional perinatal assessment: establishment of a continuous collection of indicators. Journal of Epidemiology and Public Health, 49, 583-593.
12. Decree defining the conditions under which the declarations and authorizations concerning the means and services of cryptology are subscribed, n ° 98-101 of February 24, 1998.
13. Order fixing the list of means and services of cryptology exempted from any prior formality, No. 98-206 of 23 March 1998.
14. Order fixing the list of means and services of cryptology for which the declaration replaces the authorization, n ° 98-207 of March 23, 1998.
15. Decree n ° 99-199 of March 17, 1999 defining the categories of means and services of cryptology for which the procedure of preliminary declaration is substituted for that of authorization.
16. DOUGLAS S. (1996). Cryptology, Theory and Practice, International Thomson Publishing.
17. FlSHER F., MADGE B. (1996). Data security and patient confidentiality: the manager's role. International Journal of Biomedical Computer, 43, 115-119.
18. GOUYON B., METAL P., FROMAGET J., SAGOT P., GOUYON J.-B (1999). Perinatal network of Burgundy. Technology and Health, 37, 51-56.
19. JARO M.-A. (1995). Probabilistic-linkage of large public health data files. Statistics in Medicine, 14, 491-8.
20. Law No. 98-1266 of 30 December 1998 (Article 107). Finance law for the year 1999.
21. Letter from the Institut de Veille Sanitaire, prevalence, n ° 8, July 2003.
22. MARSAULT X. (1995). Compression and encryption of multimedia data, Hermès, Paris.
23. MERLIÈRE Y. (2004). "I SNIIR-AM" communication at the Journées de Statistique May 25, 2004, Montpellier.
24. QUANTIN C, BOUZELAT H., ALLAËRT F.-A. et al. (1998). Automatic record hash coding and linkage for epidemiological follow-up data confidentiality. Methods of Information in Medicine, 37, 271-277.
25. QUANTIN C., ALLAERT F.-A., ATHis P., DUSSERRE L. (1999). Can a database be anonymous? MIE 99, Slovenia, 22-26 August 1999, 297-301.
26. QUANTIN C., ALLAERT F.-A., DUSSERRE L. (2000a). Anonymous statistical methods versus cyrptographic methods in epidemiology. International Journal of Medical Informatics, 60, 177-83.
27. QUANTIN C, ALLAERT F.-A., BOUZELAT H., RODRIGUES J.-M., TROMBERTPAVIOT B., BRUNET-LECOMTE P., GREMY F., DUSSERRE L. (2000b). The security of medical information networks: application to epidemiological studies. Journal of Epidemiology and Public Health, 48, 89-99.

28. QUANTIN C, BINQUET C, BOURQUARD K., PATTISINA R., GOUYON B., FERDYNUS C, GOUYON J.-B., ALLAERT F.-A. (2004). Which are the best identifiers for record linkage? Medical Informatics and the Internet Medicine, 29 (3-4), 221-227.
29. QUANTIN C, BINQUET C, ALLAERT F.-A., GOUYON B., PATTISINA R., TEUFF G., FERDYNUS C, GOUYON J.-B. (2005). Decision analysis for the assessment of a record linkage procedure: application to a perinatal network. Methods of Information in Medicine, 44, 72-79.

    RIVEST R.L., SHAMIR A., ADLEMAN L. (1978). A method for obtaining digital signatures and public key cryptosystems, CACM, 2, 120.
30. SUGARMAN J.-R., HOLLIDAY M., Ross A. et al. (1996). Improving American Indian cancer data in the Washington state using Indian health services and tribal records. American Cancer Society, 78 (7suppl), 1564-8.
31. SWEENEY L. (1998). Three Computational Systems for Disclosing Medical Data in the Year 1999. MEDINFO 98, IMIA, B. Cesnik, A. McCray, J.-R. Scherrer (Eds). IOS Press, Amsterdam, 1124-1129.
32. THIRION X., SAMBUC R., SAN MARCO J.-L. (1988). Epidemiology and anonymity: a new method. Journal of Pidemiology and Public Health, 36, 36-42.
33. TROUESSIN G., ALLAERT F.-A. (1997). HAY: a nominative information occultation function. MIE, 3, 196-200.
34. TROUESSIN G. "Diagnostic and Therapeutic Quality in Cancer" report: communication of multimedia information in a multidisciplinary secure network. Security of medical information in telemedicine ", study of the Ministry of Research.
35. VuiLLET-TAVERNlER S. (2000). Reflection on anonymity in the processing of health data. Medicine and Law, 40, 1-4.
36. WLLLLENBORG L.C.R.J., WALL A.G., KELLER W.J (1995). Some Methodologies! Issues in Statistical Disclosure Control. Statistics Netherlands, Department of Statistical Methods. Second Cathy Marsh Memorial Seminar, November 7, London.
37. ZIMMERMANN P. (1986). A proposed standard format for RSA cryptosystems, Boulder Software Engineering, Computer, 9, 21.

Received: 22 Sep 2019 | Revised: 13 Oct 2019 | Accepted: 15 Jan 2020