Généralitéssurlacryptographieetles codes correcteurs d'erreurs

Ce chapitre sera l'occasion d'aborder les notions de cryptographie et de codes correcteurs d'erreurs. Ces derniers sont très utilisés en cryptographie mais également en théorie de l'information et de la communication. Ce sont des codes qui ont pour objectif de permettre la transmission de l'information malgré l'ajout éventuel d'erreurs lors de cette transmission. Afin d'y parvenir, on ajoute une redondance au message à transmettre qui, lorsqu'un nombre suffisamment faible d'éléments de ce message étendu est perdu ou altéré, permettra de reconstituer le message initialement envoyé. Cette reconstitution est appelée le décodage.

1.1. Généralités sur la cryptographie

1.1.1. Définitions

La cryptographie vient du grec *kryptos* qui veut dire caché et de *graphien* qui signifie écrire. La cryptographie est donc un ensemble des techniques permettant de protéger une communication au moyen d'un code graphique secret.

Elle est l'étude des méthodes d'envoi de messages codés de telle sorte que seul le destinataire puisse le décoder. Le message qu'on veut envoyer est appelé le **texte clair** et le message codé, ou **encrypté**, est également appelé **cryptogramme**.

Le processus de conversion d'un texte clair en message codé s'appelle **chiffrement**, ou codage. Et le processus inverse est appelé **déchiffrement**, ou décodage. Pour effectuer un codage, on suit une méthode précise appelée système cryptographique, ou **cryptosystème**. Un codage se fait donc à l'aide d'un cryptosystème, et celui-ci nécessite très souvent l'utilisation d'une clé.

***Figure 1**: Schéma général de la cryptographie*

La cryptanalyse est l'étude des méthodes qui permettent de découvrir le sens d'un message codé, sans connaître le message original. Il y a plusieurs situations possibles. On peut vouloir simplement trouver le sens du message codé, sans chercher la clé de codage. Mais, en général on voudra d'abord trouver quel est le système de codage, puis la clé de codage utilisée. Lorsqu'on a trouvé tous les éléments de la méthode utilisée pour coder les messages, on dit qu'on a cassé, ou brisé, le système cryptographique en question. Plus un système est « difficile » à briser, plus il est qualifié de « sûr ».

La cryptologie est l'ensemble formé de la cryptographie et de la cryptanalyse. Elle fait partie d'un ensemble de théories et techniques liées à la transmission de l'information (théorie des ondes électromagnétiques, théorie du signal, théorie de l'information, …).

1.1.2. Principe de Kerckhoff

Dans son article *La cryptographie militaire* [7], Auguste Kerckhoff fustige le manque de précaution lors de l'utilisation des techniques cryptographiques dans les armées.

Cependant, la partie la plus connue de cet article est de *Desiderata de la cryptographie militaire*, où il énonce six conditions pour avoir un système cryptographique fiable et utilisable par les armées. De ces six points, le plus important est le second: «il faut qu'il (le système cryptographique) n'exige pas le secret, et qu'il puisse sans inconvénient tomber entre les mains de l'ennemi ».

Autrement dit, tout système cryptographique devrait avoir sa fiabilité du fait de la difficulté de trouver sa clé, et non du secret de l'algorithme. De plus, en rendant les algorithmes publics, les cryptanalystes du monde entier peuvent essayer de le casser ou l'améliorer.

1.1.3. Enjeux de la cryptographie

Dès que plusieurs entités sont impliquées dans un échange de messages sécurisés, des règles doivent déterminer l'ensemble des opérations cryptographiques à réaliser, leur séquence, afin de sécuriser la communication.

Ainsi, lorsque l'on parle de ‘‘sécuriser un échange’’, on souhaite prêter attention aux trois services suivants:

La confidentialité: seulement le destinataire autorisé doit être capable d'extraire le contenu du message de son état crypté. Par ailleurs, l'obtention de l'information à propos du contenu du message ne doit pas être possible.

L'intégrité: Par analyse de l'intégrité le destinataire peut déterminer si le message a été modifié pendant la transmission.

L'authentification: le destinataire doit avoir la capacité d'identifier le destinataire du message, aussi bien que de savoir si c'est l'auteur présumé qui a en effet envoyé le message.

1.1.4. Notion de chiffrement

Dans la cryptographie moderne, toute sécurité est basée sur la clé (ou les clés), et non dans les détails des algorithmes. Cela signifie qu'un algorithme peut être publié et analysé, mais la clé doit être protégée [8].

Ainsi, dans les systèmes cryptographiques utilisant des clés, il faut bien distinguer les deux types existants: le chiffrement symétrique et le chiffrement asymétrique.

1.1.4.1. Chiffrement symétrique

Le chiffrement symétrique (aussi appelé chiffrement à clé privée, ou à clé secrète) se base sur l'utilisation d'une clé qui doit rester secrète et qui ne doit être connue que par les personnes devant crypter et décrypter les messages. La sécurité de la méthode de chiffrement réside donc dans la difficulté à trouver cette clé.

***Figure 2**: Chiffrement symétrique*

Les algorithmes de chiffrement symétriques sont très rapides en termes de calcul. Cependant, leur principal inconvénient est le problème de la distribution de la clé entre l'expéditeur et le destinataire.

On peut distinguer deux types de chiffrements symétriques :

- Le chiffrement par flux qui se fait bit à bit sans attendre la réception entière de données. Le **RC4** (Rivest Cipher 4) est l'algorithme le plus connu pour ce type de chiffrement.
- Le chiffrement par bloc qui consiste à diviser les données en blocs de taille généralement fixe (entre 64 et 128 bits). Les blocs sont ensuite chiffrés les uns après les autres. Les algorithmes les plus répandus de ce type de chiffrement sont : le **DES** (Data Encryption Standard) et l'**AES** (Advanced Encryption Standard) qui est actuellement l'algorithme le plus utilisé et le plus sûr.

1.1.4.2. Chiffrement asymétrique

Inventé en 1977 par Diffie et Hellmann [9], le chiffrement asymétrique, aussi connu sous le nom de chiffrement à clé publique, élimine la problématique de la transmission de la clé du chiffrement symétrique.

Le principe de ce chiffrement est basé sur l'usage d'un couple de clés, l'une publique qui est connu par tout le monde et l'autre privée qui doit rester confidentielle.

***Figure 3**: Chiffrement asymétrique*

Au niveau des performances, le chiffrement asymétrique est 1000 fois plus lent que le chiffrement symétrique. Cependant, avec le chiffrement asymétrique, la distribution des clés

est grandement facilitée car l'échange des clés secrètes n'est plus nécessaire. Chaque utilisateur conserve sa clé secrète sans jamais la divulguer. Seule la clé publique devra être distribuée.

Les algorithmes les plus fréquents du chiffrement asymétrique sont : le **RSA** (Rivest Shamir Adleman) [10] qui est très utilisé dans le commerce électronique et plus généralement pour échanger des données confidentielles sur internet, **ElGamal**, **Merkle-Hellman**, etc.

1.1.5. Fonction de hachage

Les fonctions de hachage sont des fonctions qui prennent en entrée une chaine de bits de longueur arbitraire et produisent un résultat de taille fixe appelé empreinte ou hash. Elles sont en général utilisées dans les signatures numériques [11].

Les fonctions de hachage doivent être résistantes aux collisions, c'est-à-dire qu'on ne peut pas trouver facilement deux messages différents ayant la même empreinte. Bien évidemment, la taille de l'empreinte étant fixe, il existera toujours des collisions du fait qu'il y a un nombre limité d'empreintes alors que le nombre de messages initiaux est illimité.

1.2. Codes correcteurs d'erreurs

1.2.1. Théorie des codes

En 1948, dans son article intitulé ''*A mathematical theory of communication*'', Claude Shannon posa les bases de ce que l'on appelle aujourd'hui la théorie de l'information [12]. La théorie de l'information décrit les aspects les plus fondamentaux des systèmes de communication, essentiellement à l'aide de la théorie des probabilités. Elle étudie les moyens de transmettre un message depuis une source (voix, signal numérique, onde électromagnétique, ...) jusqu'à un destinataire et via un canal de transmission (liaison radio, ligne téléphonique, ...).

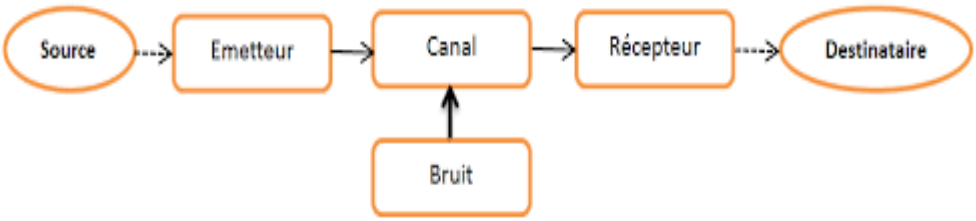

***Figure 4**: Synoptique d'un système de communication*

Cependant, lorsqu'une information est transmise à travers un canal, celui-ci génère des erreurs ou des effacements. Le canal est alors bruité. Ce problème sera résolu en 1950, lorsque Richard W. Hamming, dans son article ''*Error detecting and error correcting codes*'' introduisit la théorie des codes [13]. Hamming proposa ainsi les premiers codes correcteurs d'erreurs. Ces derniers sont un outil qui permet d'assurer la transmission fiable d'informations à travers un canal. L'idée étant de rajouter de la redondance au message transmis, de sorte que cet excèdent d'information aide à retrouver le message initial. Une grande famille des codes correcteurs d'erreurs est constituée des codes par blocs. Pour ces codes, l'information est d'abord coupée en blocs de taille constante et chaque bloc est transmis indépendamment des autres, avec une redondance qui lui est propre. La plus grande sous-famille de ces codes rassemble ce que l'on appelle les codes linéaires.

1.2.2. Canal binaire symétrique

La modélisation des canaux est un enjeu majeur pour la théorie de l'information afin de comprendre les propriétés du bruit qu'ils vont produire et de l'estimer. En effet, le choix de l'algorithme de décodage et le bon fonctionnement de celui-ci dépendent du modèle du canal utilisé.

Dans notre manuscrit, nous allons considérer le modèle du canal binaire symétrique. C'est un canal binaire caractérisé par la probabilité d'erreur p qu'au cours de la transmission un bit (0 ou 1) soit modifié en son opposé. Ces modifications se produisent indépendamment sur chacun des bits transmis. Nous illustrons ce canal par la figure ci-dessous.

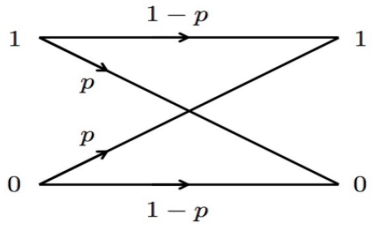

***Figure 5**: Canal binaire symétrique de probabilité d'erreur p*

1.2.3. Codes linéaires

1.2.3.1. Définition et exemple

Soit $\mathbb{F}$ un corps fini et $n > 0$.

Un **code linéaire binaire** $\mathcal{C}$ de longueur n et de dimension k est un sous-espace vectoriel engendré par k vecteurs libres de $\mathbb{F}_2^n$ [15]. Le nombre de mots d'un tel code est donc 2^k. On notera ce code un $[n, k]$-code linéaire.

En d'autres termes, un code linéaire est un ensemble de vecteurs de longueurs n, qui est stable par addition et par multiplication par un scalaire, dont une base contient k vecteurs linéairement indépendants.

Exemple : Le code {0000, 1101, 0110, 1011} est un $[n, k]$-code linéaire binaire de longueur 4 et de dimension 2.

Le poids de Hamming de v, noté $wt(v)$ est le nombre de ses coefficients non nuls.

La distance de Hamming entre v et v' est définie par :

$$d(v, v') = wt(v' - v).$$

La distance minimale d'un code est la plus petite distance entre deux mots distincts du code.

Parallèlement, nous définissons le support d'un vecteur comme les indices des coordonnées non nulles de celui-ci.

Le support de $v \in \mathbb{F}_2^n$ est défini comme :

$$Supp(v) = \{j : v_j = 1\} \subseteq \{1, \dots, n\}$$

Le poids de Hamming d'un vecteur $v \in \mathbb{F}_2^n$ se définit de manière équivalent comme :

$$wt(v) = |Supp(v)|.$$

Un code linéaire, étant un sous-espace vectoriel, peut être décrit grâce à une base de vecteurs représentée sous forme matricielle. Ainsi, on distingue la matrice génératrice et la matrice de contrôle de parité.

1.2.3.2. Matrice génératrice

Soit $\mathcal{C}$ un [n, k]-code linéaire binaire. On appelle **matrice génératrice** de $\mathcal{C}$, notée G, toute matrice de taille $k \times n$ dont les lignes forment une base de $\mathcal{C}$. On a alors :

$$\mathcal{C} = \{\vartheta \mathrm{G} \mid \vartheta \in \mathbb{F}_2^k \}$$

La matrice G définie une bijection de $\mathbb{F}_2^k \longrightarrow \mathcal{C}$ par $\vartheta \longmapsto \vartheta G$ et ainsi on représente 2^k messages distincts par des mots du code $\mathcal{C}$. Chaque mot de longueur k est codé par un mot du code $\mathcal{C}$ de longueur n.

Une matrice génératrice G d'un code $\mathcal{C}$ n'est pas unique. Puisque les k colonnes de G sont linéairement indépendantes, en effectuant des opérations élémentaires sur les lignes, G peut être transformée en $G^* = (I_k|A)$, I_k est la matrice identité d'ordre k et A une matrice $k \times (n - k)$.

G^* est appelée matrice génératrice canonique de $\mathcal{C}$.

Ainsi, lorsqu'un code linéaire $\mathcal{C}$ possède une matrice génératrice de la forme $G^* = (I_k|A)$, on dit que le code $\mathcal{C}$ est **systématique** [16].

1.2.3.3. Code dual

Soit $\mathcal{C}$ un [n, k]-code linéaire binaire.

Le code dual (ou code orthogonal), noté $\mathcal{C}^{\perp}$, est l'espace-vectoriel orthogonal de $\mathcal{C}$ pour le produit scalaire sur $\mathbb{F}_2^n$ [18]. On note :

$$\mathcal{C}^{\perp} = \{x \in \mathbb{F}_2^n \mid x.y = 0, \forall\, y \in \mathcal{C}\}$$

1.2.3.4. Matrice de contrôle de parité

Soit $\mathcal{C}$ un [n, k]-code linéaire binaire.

On appelle matrice de contrôle de parité de $\mathcal{C}$, toute matrice génératrice du code dual $\mathcal{C}^{\perp}$, notée H. On a :

$$\mathcal{C} = \{x \in \mathbb{F}_2^n \mid x.H^t = 0\}$$

La matrice de contrôle de parité, comme son nom l'indique, permet le contrôle d'erreurs. Ainsi, pour toute matrice génératrice G, il existe une matrice de contrôle H, telle que les lignes de H soient orthogonales aux lignes de la matrice G. Autrement dit, $GH^t = 0$.

Comme pour la matrice génératrice G, on peut écrire la matrice de contrôle de parité sous forme systématique. Dans ce cas on a :

$$H = (-A^t | I_{n-k})$$

1.2.3.5. Fonction syndrome

Soit H une matrice de contrôle de parité d'un code $\mathcal{C}$, le syndrome $s \in \mathbb{F}_2^{n-k}$ de tout vecteur $x \in \mathbb{F}_2^n$ est définit comme suit :

$$s = H.x^T$$

Remarque : En multipliant un mot de code du code $\mathcal{C}$ par sa matrice de contrôle de parité H, on obtient le vecteur, c'est-à-dire : $\forall\, c \in \mathcal{C},\ Hc^T = 0$.

1.2.4. Quelques exemples de codes linéaires

1.2.4.1. Code de Hamming

Un code de Hamming est un code correcteur linéaire. Il est utilisé dans les transmissions de données car il permet de détecter et de corriger une erreur survenue dans un bloc transmis [17].

Soit m un entier positif supérieur ou égal à 2. **Un code de Hamming binaire**, dit de paramètre r, est un code linéaire défini sur $\mathbb{F}_2$, de longueur $n = 2^r - 1$, de dimension $k = 2^r - 1 - r = n - r$ et de distance minimale $d = 3$, Les colonnes d'une matrice de contrôle d'un code de Hamming binaire sont tous les vecteurs non nuls de $\mathbb{F}_2^r$ [18].

Le code de Hamming est un code parfait, ce qui signifie que pour une longueur de code donnée, il n'existe pas d'autre code plus compact ayant la même capacité de correction. En ce sens, son rendement est maximal.

1.2.4.2. Codes de Goppa classique

En 1970, le mathématicien russe, Valery Denisovich Goppa découvre la relation entre la géométrie algébrique et les codes. Ce qui a pour conséquence la naissance des codes dits de Goppa classique, définis sur les corps finis [20].

Définition : Soient $\mathcal{L} = \{\alpha_0, \dots, \alpha_{n-1}\} \in \mathbb{F}_q^n$ et $G(X) \in \mathbb{F}_q[X]$ un polynôme de degré t tels que $\forall\, i \in \{0,2, \dots, n-1\}$, $g(\alpha_i) \neq 0$. Le code de Goppa de longueur n défini sur $\mathbb{F}_q$ est donné par :

$$\Gamma_q(\mathcal{L}, G) = \{C = (c_0, c_2, \dots, c_{n-1}) \in \mathbb{F}_q^n : \sum_{i=0}^{n} \frac{c_i}{X - \alpha_i} \equiv 0 \ mod\ G(X)\}$$

On dit alors que $\mathcal{L}$ est le support et G le polynôme de Goppa de $\Gamma_q(\mathcal{L}, G)$.

Grâce à la précédente équation, on peut retrouver la matrice de contrôle du code de Goppa à partir de la matrice suivante :

$$H = \begin{pmatrix} \frac{1}{g(\alpha_0)} & \cdots & \frac{1}{g(\alpha_{n-1})} \\ \vdots & \vdots & \vdots \\ \frac{\alpha_0^{t-1}}{g(\alpha_0)} & \cdots & \frac{\alpha_{n-1}^{t-1}}{g(\alpha_{n-1})} \end{pmatrix}$$

Nous verrons dans le chapitre suivant que les codes de Goppa sont très utilisés en cryptographie, notamment dans le cryptosystème de McEliece.

1.2.4.3. Codes cycliques

Les codes cycliques constituent l'une des classes les plus importantes des codes linéaires, et sont les plus utilisés en pratique. En effet, ils présentent de nombreux avantages : leur mise en œuvre (codage/décodage) est facile, ils offrent une gamme étendue de codes (comme les codes quasi-cycliques, qu'on abordera dans cette partie), et enfin permettent de corriger différent types d'erreurs.

a. Définition

Un code $\mathcal{C}$ de longueur n est dit cyclique s'il est linéaire et s'il vérifie la propriété suivante :

$$(c_0, c_1, \dots, c_{n-1}) \in \mathcal{C} \iff (c_{n-1}, c_0, c_1, \dots c_{n-2}) \in \mathcal{C}$$

En d'autres termes, dans un code cyclique, tout décalage cyclique d'un mot de code forme encore un mot de code [15].

Exemple : Le code $\mathcal{C}$ = {000, 110, 011, 101} est un code binaire cyclique.

b. Représentation polynomiale

Pour faciliter l'écriture et afin d'étudier les propriétés algébriques de ces codes, il est plus commode d'écrire les mots d'un code cyclique sous forme polynomiale.

Soit $\mathcal{C}$ un code linéaire de longueur n, à chaque mot c de $\mathcal{C}$, on associe un polynôme $c(x)$ de $\mathbb{F}_2[X]$ défini par :

$$c = (c_o, c_1, \dots, c_{n-1}) \longrightarrow c(X) = c_0 + c_1 X + c_2 X^2 + \cdots + c_{n-1} X^{n-1}$$

Ainsi, l'action du décalage circulaire sur un mot de code revient à la multiplication par X modulo $X^n - 1$ sur le polynôme correspondant :

$$(c_{n-1}, c_0, \dots, c_{n-2}) \longrightarrow c(X) = c_{n-1} + c_0 X + \cdots + c_{n-2} X^{n-1}$$

Avec,

$$c_{n-1} + c_0X + \cdots + c_{n-2}X^{n-1} = X.c(X)[\text{modulo } (X^n - 1)]$$

Autrement dit, le polynôme $c_{n-1} + c_0X + \cdots + c_{n-2}X^{n-1}$ est égal au reste du polynôme $X.c(X)$ dans la division euclidienne par $X^n - 1$ [19].

c. Matrice circulante et quasi-circulante

Une matrice carrée est dite circulante si elle est construite de sorte que ses lignes soient les décalages cycliques successifs de la première ligne [15].

Une telle matrice s'écrit comme suit :

$$M = \begin{bmatrix} a_0 & a_1 & a_2 & \cdots & a_n \\ a_n & a_0 & a_1 & \cdots & a_{n-1} \\ a_{n-1} & a_n & a_0 & \cdots & a_{n-2} \\ & & & \ddots & \\ a_1 & a_2 & a_3 & \cdots & a_0 \end{bmatrix}$$

La matrice circulante est donc entièrement définie par la donnée d'une seule de ses lignes.

Remarque : Un code cyclique admet au moins une matrice génératrice (ou de contrôle de parité) circulante.

Une matrice est dite quasi-circulante lorsqu'elle est formée de blocs circulants. Une telle matrice s'écrit comme suit :

$$M = \begin{bmatrix} a_0 & a_1 & a_2 & b_0 & b_1 & b_2 \\ a_2 & a_0 & a_1 & b_2 & b_0 & b_1 \\ a_1 & a_2 & a_0 & b_1 & b_2 & b_0 \end{bmatrix}$$

On remarque que cette matrice est formée de deux blocs circulants. On a donc :

$M = [A \quad | \quad B]$, avec

$$A = \begin{bmatrix} a_0 & a_1 & a_2 \\ a_2 & a_0 & a_1 \\ a_1 & a_2 & a_0 \end{bmatrix} \quad et \quad B = \begin{bmatrix} b_0 & b_1 & b_2 \\ b_2 & b_0 & b_1 \\ b_1 & b_2 & b_0 \end{bmatrix}$$

d. Codes quasi-cycliques

Au sein des codes correcteurs d'erreurs et de la théorie de l'information, les codes quasi-cycliques tiennent une place particulière. Ces codes sont une généralisation des codes cycliques et possèdent en plus d'excellents paramètres : ils ont une grande capacité de correction.

Définition : Un code est dit quasi-cyclique lorsqu'il admet une matrice génératrice (ou de contrôle de parité) quasi-circulante.

Remarque : Un code quasi-cyclique admet au moins une matrice génératrice sous forme systématique et quasi-circulante.

Dans les chapitres suivant de notre manuscrit, on utilisera les codes quasi-cycliques qu'on va appliquer à la cryptographie.

1.2.4.4. Autres codes

Il existe beaucoup d'autres codes linéaires (les codes **BCH**, les codes **Reed-Muller**, les codes **Reed-Solomon**, les codes **alternants**, les codes **Golay**, etc.) toutes présentant leurs intérêts, soit du point de vue de la capacité de correction d'erreurs, soit de la simplicité du codage ou du décodage.

On ne fait ici que mentionner leur existence car leur connaissance n'est pas nécessaire à la compréhension de notre manuscrit. Il n'est pas en revanche exclu que des applications cryptographiques de la théorie algébrique des codes puissent passer par l'utilisation de ces codes.