Chapitre V

Cryptographie classique

1 Introduction

Cette première partie du cours sur la cryptographie a pour objectif de :

- définir les concepts de base de la cryptographie,
- présenter quelques méthodes de cryptographie classique,
- montrer comment les messages cryptés issus de ces méthodes peuvent être cassés,
- introduire au problème de la vulnérabilité des méthodes de cryptage,
- tenter de définir ce que l'on peut attendre d'un méthode de cryptage.

L'objectif de la cryptographie est de permettre de conserver une information confidentielle

- soit lorsqu'elle est communiquée d'un individu à un autre à travers un canal de communication peu ou pas sûr.
- soit lorsqu'elle est conservée sur un support de stockage ou dans un lieu considéré comme peu ou pas sûr.

Définition 20 (Cryptographie)**.** L'art et la science de garder le secret des messages.

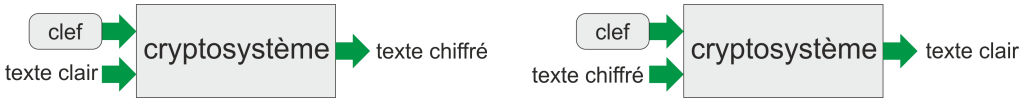
On donne alors les définitions suivantes :

Définition 21 (Chiffrement (ou encryption))**.** Processus de transformation d'un texte en clair en un texte chiffré (ou cryptogramme) de manière à le rendre incompréhensible.

Définition 22 (Déchiffrement (ou décryptage))**.** Processus de reconstruction du texte en clair à partir du texte chiffré.

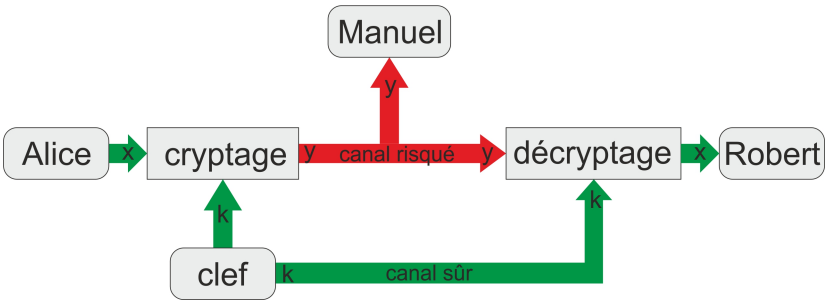
Définition 23 (Cryptosystème)**.** Système implémentant les processus de chiffrement et déchiffrement.

Définition 24 (Clef de chiffrement)**.** Paramètre d'un cryptosystème permettant de changer le résultat du chiffrement, et nécessaire pour effectuer le déchiffrement.



Si le cryptosystème n'a pas besoin de clef, il suffit de disposer du cryptosystème (ou le connaître) pour déchiffrer un cryptogramme.

Définition 25 (Cryptanalyse)**.** Art de décrypter les messages sans connaître le cryptosystème ou la clef.



Dans le modèle de communication confidentielle ci-dessus :

- Alice et Robert choisissent un cryptosystème C et une clé k de cryptage (ou Alice transmet à Robert sa clef k de cryptage par un moyen sûr).
- Pour chaque message qu'Alice envoie, elle utilise son cryptosystème C et sa clé k qui lui permet de construire un message crypté y à partir de x.
- Puis elle transmet à Robert le message crypté y sur un canal peu sûr, qui pourrait être surveillé par un individu hostile (nommé ici Manuel).
- Robert décrypte le message y en utilisant le cryptosystème C et la clé k, pour obtenir le message original x.

2 Formalisation

2.1 Cryptosystème

Définition 26 (Cryptosystème (définition formelle))**.** Un cryptosystème est un quintuplet $(\mathcal{P}, \mathcal{C}, \mathcal{K}, \mathcal{E}, \mathcal{D})$ où :

- $\mathcal{P}$ est l'ensemble des messages en clair possibles,
- C est l'ensemble des messages codés possibles,
- $\mathcal{K}$ est l'ensemble des clefs possibles,
- $\mathcal{E}$ est l'ensemble des règles de chiffrement de $\mathcal{P} \to C$,
- $\mathcal{D}$ est l'ensemble des règles de déchiffrement de $C \to \mathcal{P}$

tel que pour toute clef $k \in \mathcal{K}$, il existe une règle de chiffrement $e_k \in \mathcal{E}$ et une règle de déchiffrement $d_k \in \mathcal{D}$ vérifiant $d_k(e_k(x)) = x$ pour tout $x \in \mathcal{P}$.

Clairement, on a :

- si $x \in \mathcal{P}$ représente un message à chiffrer, et $k \in \mathcal{K}$ la clef de chiffrage, alors $e_k(x) \in C$ est le message chiffré.
- si $y \in C$ représente un message à déchiffrer, et $k \in \mathcal{K}$ la clef de chiffrage, alors $d_k(x) \in \mathcal{D}$ est le message déchiffré.
- toute fonction de e_k est une injection (*i.e.* $e_k(x_1) = e_k(x_2) \Rightarrow x_1 = x_2$).
- si $\mathcal{P} = C$, alors toute fonction de chiffrement est une permutation.

2.2 Symboles

Dans les exemples qui suivront nous considérerons uniquement le codage des messages à partir de symbole uniquement tirés des lettres majuscules de l'alphabet (sans espace, ponctuation, minuscules, accents, ...).
Par exemple, si Alice veut envoyer le message :

"Rencontre au parc Monceau à minuit",

le texte en clair à l'entrée du cryptosystème sera :

"RENCONTREAUPARCMONCEAUAMINUIT".

La charge de replacer les blancs et les accents est laissé à la charge de Robert, une fois le message décodé.
L'ensemble des caractères utilisés dans les textes des messages est donc l'ensemble des 26 symboles :

$\{A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, W, X, Y, Z\}$

Cette restriction de l'ensemble des symboles à coder est tout d'abord dans un but pratique du cours (= limiter l'espace de codage).

2.3 Arithmétique modulo

On note $\mathbb{Z}_n$ l'ensemble $\{0, 1, \ldots, n-1\}$.

Arithmétique modulo

L'addition et la multiplication dans $\mathbb{Z}_n$ sont définis exactement comme l'addition et la multiplication normale, à l'exception que les résultats sont réduits modulo n.

Exemple

Pour calculer 11×13 dans $\mathbb{Z}_{16}$. $11 \times 13 = 143$. Or, $143 \bmod 16 = 15$. Donc, $11 \times 13 = 15$ dans $\mathbb{Z}_{16}$.

Espace des symboles

Par la suite, pour coder un texte chiffré, $\{0, \ldots, 25\}$ étant homéomorphe à $\{A, B, \ldots, Z\}$, $\mathbb{Z}_{26}$ sera utilisé pour signifier l'ensemble des symboles de l'alphabet (A=0, B=1, ...).

Espace des messages

On note $\mathbb{Z}_{26}^*$ l'ensemble des chaînes dont chaque caractère est dans $\mathbb{Z}_{26}$.
i.e. l'union des espaces produits $\mathbb{Z}_{26}^* = \cup_{i=1}^{\infty} \mathbb{Z}_{26}^i$ où $\mathbb{Z}_{26}^i = \underbrace{\mathbb{Z}_{26} \times \ldots \times \mathbb{Z}_{26}}_{i \text{ fois}}$.
donc $\mathbb{Z}_{26}^*$ est homéomorphe à l'ensemble des messages possibles écrits avec des lettres de l'alphabet.

3 Chiffres classiques

3.1 Typologie des chiffres

Pour les chiffres classiques, la typologie est la suivante :

- **chiffre de transposition** (ou de permutation) : l'ordre des lettres est changé.
- **chiffre de substitution** : les lettres ou les mots sont remplacés par d'autres symboles. Il y a plusieurs sous-familles :
  - ◇ les chiffres **monoalphabétiques** : chaque lettre est remplacé par une lettre ou un autre symbole.
  - ◇ les chiffres **polyalphabétiques** : chaque lettre est remplacé par un ou plusieurs symboles.
    pour les chiffres **homophoniques**, le nombre de symboles remplaçant une lettre est proportionnel à sa fréquence d'apparition.
  - ◇ les chiffres **polygrammiques** (ou polygraphique) : les lettres sont chiffrés par paquets de plusieurs lettres (deux ou trois).
  - ◇ les chiffres **tomogrammiques** (ou par fraction de lettres) : chaque lettre est représentée par des groupes de symboles, qui sont ensuite chiffrés séparément par substitution ou transposition.
    Ceux-ci n'ont eu pour l'instant que peu d'intérêt car ils produisent des chiffres beaucoup plus longs que le texte clair.

3.2 Monoalphabétiques

Dans le cas des chiffres mono-alphabétiques, chaque lettre étant remplacé par un ou plusieurs symboles, la chaîne peut être codée symbole par symbole.

Définition 27 (Chiffre mono-alphabétique d'une chaîne)**.** Un chiffre mono-alphabétique est un cryptosystème qui chiffre/déchiffre ses symboles l'un après l'autre, à savoir :

Chiffrement d'un chaîne $X = x_1x_2 \dots x_n \in \mathcal{P}$,

$e_k(X) = e_k(x_1)e_k(x_2) \dots e_k(x_n) = y_1y_2 \dots y_n = Y$.

Déchiffrement d'un code $Y = y_1y_2 \dots y_n \in C$,

$d_k(Y) = d_k(y_1)d_k(y_2) \dots d_k(y_n) = x_1x_2 \dots x_n = X$.

Donc, dans les chiffrages mono-alphabétiques que nous aborderons :

- nous ne donnerons la règle de chiffrage que d'un seul symbole,
- le chiffrage d'une chaîne de symboles est obtenu avec la règle ci-dessus.

Conséquence

Dans un chiffre monoalphabétique, chaque symbole (ou groupe de symboles) est codé par un chiffre unique.

3.2.1 Chiffre mono-alphabétique par décalage

Définition 28 (Chiffre par décalage)**.** Soit $\mathcal{K} = \mathbb{Z}_{26}$ et $\mathcal{P} = C = \mathbb{Z}_{26}^*$.
Un chiffrage par décalage de clé $k \in \mathcal{K}$ est défini par les règles suivantes :

- pour $x \in \mathbb{Z}_{26}$, $e_k(x) = (x + k) \bmod 26$.
- pour $y \in \mathbb{Z}_{26}$, $d_k(y) = (y - k) \bmod 26$.

Exemples

pour k=12, la lettre R (=17) est chiffrée comme :

17+k mod 26 = 29 mod 26 = 3, donc D.

pour k=5, la lettre C (=2) est déchiffrée comme :

2-k mod 26 = -3 mod 26 = 23, donc X.

Remarques :

- Le chiffrage de César est un chiffrage par décalage pour lequel $k = 3$.
- Noter que pour $k = 0$, il n'y a pas de décalage, donc de chiffrage.
- Ce chiffre ne paraît pas fiable en raison du petit nombre de codes possibles ($\#\mathcal{K} - 1 = 25$).

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z

Exemple 1 (par décalage dans l'alphabet)

X ="RENCONTREAUSQUAREAMINUIT" et $k = 3$

Le chiffrement s'obtient avec un décalage de 3 symboles à droite.

$e_3(X)$ ="UHQFRQWUHDXVTXDUHDPLQXLW"

Le déchiffrement s'obtient avec un décalage de 3 symboles à gauche.

Exemple 2 (par calcul arithmétique)

X ="RENCONTREAUSQUAREAMINUIT" et $k = 15$

x_i	R	E	N	C	O	N	T	R	E	A	U	S	Q	U	A	R	E	A	M	I	N	U	I	T
x_i	17	4	13	2	14	13	19	17	4	0	20	18	16	20	0	17	4	0	12	8	13	20	8	19
x_i+k	32	19	28	17	29	28	34	32	19	15	35	33	31	35	15	32	19	15	27	23	28	35	23	34
%26	6	19	2	17	3	2	8	6	19	15	9	7	5	9	15	6	19	15	1	23	2	9	23	8
y_i	G	T	C	R	D	C	I	G	T	P	J	H	F	J	P	G	T	P	B	X	C	J	X	I

$e_{15}(X)$ ="GTCRDCIGTPJHFJPGTPBXCJXI"

Le déchiffrement s'effectuer en retranchant k, toujours en arithmétique modulaire.

Exercice 32: Chiffre par décalage

1. Coder le message suivant "Qui pense peu se trompe beaucoup" avec $k = 7$ en utilisant un chiffre à décalage.
2. Décoder le message "FJXCTSDJITEPHPRFJXTGIETJ" codé avec un chiffre à décalage et la clef $k = 15$.

3.2.2 Chiffre mono-alphabétique par substitution

Définition 29 (permutation)**.** Une permutation π sur un ensemble discret E est une bijection de E sur E.

i.e. une permutation est une fonction qui mélange les éléments d'un ensemble.

Remarque : On note π^{-1} la permutation inverse (existe car π est une bijection).

Exemple

Soit $\mathbb{Z}_3 = \{0, 1, 2\}$.
Un exemple de permutation $\pi : \mathbb{Z}_3 \to \mathbb{Z}_3$ est :
$\pi(0) = 2, \pi(1) = 0, \pi(2) = 1$ (*i.e.* $\pi(\{0, 1, 2\}) = \{2, 0, 1\}$).
$\pi^{-1}(\{2, 0, 1\}) = \{0, 1, 2\}$.

Proposition 18 (nombre de permutations)**.** *Si E est un ensemble dont le cardinal est #E = n, alors le nombre de permutations possibles est n*!

Démonstration:
Dans un ensemble à n éléments, il y n choix pour placer le premier élément, $n - 1$ pour placer le deuxième, $n - 2$ pour placer le troisième, ..., 1 pour placer le dernier. En conséquence, il y a $n!$ façons différentes de construire une permutation. □

Définition 30 (Chiffre mono-alphabétique par substitution (antiquité))**.** Soit $\mathcal{P} = C = \mathbb{Z}_{26}^*$. Soit $\mathcal{K}$ l'ensemble des permutations sur $\mathbb{Z}_{26}$.

Un chiffrage par substitution de clef $\pi \in \mathcal{K}$ est défini par les règles suivantes de codage par symbole :

- pour $x \in \mathbb{Z}_{26}$, $e_k(x) = \pi(x)$.
- pour $y \in \mathbb{Z}_{26}$, $d_k(y) = \pi^{-1}(y)$.

Exemple de fonction de permutation :

x	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
π	24	22	3	15	21	14	9	20	13	7	19	5	4	10	6	16	11	12	2	8	23	25	17	1	0	18

ce qui revient à effectuer les permutations de lettres suivantes :

x	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
π	Y	W	D	P	V	O	J	U	N	H	T	F	E	K	G	Q	L	M	C	I	X	Z	R	B	A	S

chiffrage : chercher le x dans la première ligne, chiffrer avec $\pi(x)$ dans la seconde.

déchiffrement : chercher le $\pi(x)$ dans la seconde ligne, déchiffrer avec x dans la première.

Exemple de chiffrement

Chiffrement de X ="RENCONTREAUSQUAREAMINUIT".
On utilise la permutation π ci-dessus.
On obtient Y ="MVKDGKIMVYXCLXYMVYENKXNI".
Le déchiffrement est trivial (opération inverse).

Remarque

Par rapport au code par décalage, un code par substitution a un espace de clefs de taille :

$\#\mathcal{K} = \#\mathbb{Z}_{26}! = 26! = 403.291.461.126.605.635.584.000.000 > 4.10^{26}$

En supposant que l'on teste 1.000 milliards de permutations à la seconde, 12,7 millions d'années sont nécessaires pour vérifier toutes les permutations.
Heureusement, en moyenne, il ne faudra en tester que la moitié ...
Conséquence : une attaque par force brute est vaine (et stupide).

Construction de π^{-1} : utiliser le fait que $\pi^{-1}(\pi(x)) = x$.

Mais comment transmettre simplement une substitution ?

1. on construit un carré en mettant les lettres de la clef sans répétition sur la première ligne, et en le complétant avec le reste de l'alphabet (sans les lettres déjà utilisées) sur les lignes suivantes.
2. on trie les colonnes du tableau par ordre alphabétique.
3. on aligne les colonnes pour former les substitutions.

Prenons comme clef le mot BATEAU.

B	A	T	E	U
C	D	F	G	H
I	J	K	L	M
N	O	P	Q	R
S	V	W	X	Y
Z				

A	**B**	**E**	**T**	**U**
D	C	G	F	H
J	I	L	K	M
O	N	Q	P	R
V	S	X	W	Y
	Z			

x	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
π	**A**	D	J	O	V	**B**	C	I	N	S	Z	**E**	G	L	Q	X	**T**	F	K	P	W	**U**	H	M	R	Y

De nombreuses variations de la construction de ces permutations existent.

Exercice 33: Chiffre par substitution

1. Créer la table de substitution associée au mot-clef PASTEQUE.
2. Coder le message "urgent/fort de la Pompelle/pénurie munitions".
3. Décoder le message `KOHTLEATTADFEWEQAPHAQEXXPHEX`.

3.3 Polyalphabétiques

Les chiffres polyalphabétiques peuvent utiliser plusieurs symboles différents pour représenter le même symbole.
Le plus célèbre et le plus utilisé (avec ses variations) est le chiffre de Vigenère.

Définition 31 (Chiffre de Vigenère (XVI$^{\text{ème}}$ siècle))**.** Soit $\mathcal{P} = C = \mathbb{Z}_{26}^*$. Soit $\mathcal{K}$ l'ensemble des clefs sur $\mathbb{Z}_{26}^*$.
Un chiffre de Vigenère de clé $k = k_0 k_1 \dots k_{p-1}$ (de longueur p) sur un mot $m = m_0 m_1 \dots m_{q-1}$ (de longueur q) est défini par les règles suivantes :

- $e_k(m_i) = (m_i + k_{i \bmod p}) \bmod 26$
- $d_k(y_i) = (y_i - k_{i \bmod p}) \bmod 26$

Exemple

Si la clef est k ="MACLEF" $= \{12, 0, 2, 11, 4, 5\}$ et le message à coder est m ="MONMESSAGEACODER", alors le chiffre de Vigenère est :

m	M	O	N	M	E	S	S	A	G	E	A	C	O	D	E	R
k	M	A	C	L	E	F	M	A	C	L	E	F	M	A	C	L
m_i	12	14	13	12	4	18	18	0	6	4	0	2	14	3	4	17
$k_{i \bmod p}$	12	0	2	11	4	5	12	0	2	11	4	5	12	0	2	11
$e_k(c_i)$	24	14	15	23	8	23	4	0	8	15	4	7	0	3	6	2
	Y	O	P	X	I	X	E	A	I	P	E	H	A	D	G	C

Exercice 34: Chiffre de Vigenère

1. Coder le message "Qui ne doute pas acquiert peu" avec la clef "PASTEQUE".
2. Décoder le message `AEKVSDPMRTAHRIMSCTVXWFLMHOFL` avec la même clef.

On remarquera que dans ce message :

- une lettre n'est jamais codée deux fois de la même façon,
- inversement, deux lettres différentes peuvent donner la même lettre dans le chiffre.

Cette caractéristique rend la cryptanalyse de ce code plus difficile.
Différentes variations simples de ce code existent :

- le code de Beaufort $e_k(m_i) = k_{i \bmod p} - m_i$
- la variante Allemande $e_k(m_i) = m_i - k_{i \bmod p}$
- la variante de Rozier est en fait un code de Vigenère donc la clef est calculée est calculée comme $k'_i = k_{p-1-i} - k_i + 1$.

La machine électromécanique Enigma, utilisée pour chiffrer les transmissions militaires allemandes pendant la seconde guerre mondiale, utilise également ce principe (en l'automatisant) en générant des clefs de plusieurs centaines de millions de lettres.

La clef était déterminée par le choix, l'ordre et la position des rotors, et de la connexion de 10 cables pour relier des lettres entres-elles, générant un total de 1.6×10^{20} chiffres possibles.

3.4 Polygrammiques

Le principal inconvénient des chiffres alphabétiques est que ceux-ci codent les caractères un à un, ce qui facilite la cryptanalyse en raison de la faiblesse du chiffre ou de la clef employée.

Les chiffres polygrammiques résolvent ce problème en chiffrant les symboles par paquets de n.

Cette approche est intéressant à double titre :

- la distribution de probabilité des n-grams est beaucoup plus plate que celle des 1-grams.
  Il est donc beaucoup plus difficile de reconnaitre des fragments.
- le nombre de n-grams devient très important pour $n \geq 4$.

L'un des codes polygrammiques les plus intéressants consiste à utiliser l'algèbre matricielle dans $\mathbb{Z}_{26}$ afin de combiner les symboles par paquets de n :

Définition 32 (Chiffre de Hill (XX$^{\text{ème}}$ siècle))**.** Soit $\mathcal{P} = C = \mathbb{Z}_{26}^*$.
Soit $m \geq 2$ et $\mathcal{K}$ l'ensemble des matrices de taille $m \times m$ inversibles sur $\mathbb{Z}_{26}$.
Un chiffre de Hill pour une clé $K \in \mathcal{K}$ est défini par les règles suivantes :
Chiffrement : soit $M = M_1 M_2 \dots M_n$ un texte clair
où M_i représente le $i^{\text{ème}}$ paquets de m symboles.
$e_k(M) = e_k(M_1) \dots e_k(M_n)$ et $\forall i,\ e_k(M_i) = K.M_i \bmod 26$.
Déchiffrement : Soit $Y = Y_1 Y_2 \dots Y_n$ un texte chiffré
où Y_i représente le $i^{\text{ème}}$ paquets de m symboles chiffrés.
$d_k(Y) = d_k(Y_1) \dots d_k(Y_n)$ et $\forall i,\ d_k(Y_i) = K^{-1}.Y_i \bmod 26$.

Exemple

On veut chiffrer le message M ="MONMESSAGEACODER" avec un chiffre de Hill utilisant des blocs de 2 lettres.
On prend la clef $K = \begin{bmatrix} 11 & 3 \\ 8 & 7 \end{bmatrix}$.
Pour le premier bloc : M_1="MO" $= \begin{bmatrix} 12 \\ 14 \end{bmatrix}$.
Donc, $Y_1 = K.M_1 = \begin{bmatrix} 11 & 3 \\ 8 & 7 \end{bmatrix} . \begin{bmatrix} 12 \\ 14 \end{bmatrix} = \begin{bmatrix} 174 \\ 194 \end{bmatrix} = \begin{bmatrix} 18 \\ 12 \end{bmatrix}$ = "SM"
Pour les autres blocs :

M	MO	NM	ES	SA	GE	AC	OD	ER
M_i	$\begin{bmatrix}12\\14\end{bmatrix}$	$\begin{bmatrix}13\\12\end{bmatrix}$	$\begin{bmatrix}4\\18\end{bmatrix}$	$\begin{bmatrix}18\\0\end{bmatrix}$	$\begin{bmatrix}6\\4\end{bmatrix}$	$\begin{bmatrix}0\\2\end{bmatrix}$	$\begin{bmatrix}14\\3\end{bmatrix}$	$\begin{bmatrix}4\\17\end{bmatrix}$
$K.M_i$	$\begin{bmatrix}18\\12\end{bmatrix}$	$\begin{bmatrix}23\\6\end{bmatrix}$	$\begin{bmatrix}20\\2\end{bmatrix}$	$\begin{bmatrix}16\\14\end{bmatrix}$	$\begin{bmatrix}0\\24\end{bmatrix}$	$\begin{bmatrix}6\\14\end{bmatrix}$	$\begin{bmatrix}7\\3\end{bmatrix}$	$\begin{bmatrix}17\\21\end{bmatrix}$
Y_i	SM	XG	UC	QO	AY	GO	HD	RV

Donc Y ="SMXGUCQOAYGOHDRV"
K est bien inversible.
La clef inverse est : $K^{-1} = \begin{bmatrix}7 & 23\\18 & 11\end{bmatrix}$.
En effet, $K.K^{-1} = \begin{bmatrix}11 & 3\\8 & 7\end{bmatrix}.\begin{bmatrix}7 & 23\\18 & 11\end{bmatrix} = \begin{bmatrix}131 & 286\\182 & 261\end{bmatrix} = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$.
Le premier bloc à décoder est Y_1 = "SM" $= \begin{bmatrix}18\\12\end{bmatrix}$
$M_1 = K^{-1}.Y_1 = \begin{bmatrix}7 & 23\\18 & 11\end{bmatrix}.\begin{bmatrix}18\\12\end{bmatrix} = \begin{bmatrix}402\\456\end{bmatrix} = \begin{bmatrix}12\\14\end{bmatrix}$ = "MO"
etc ...

Exercice 35: Chiffre de Hill

1. Soit le message "Veni, vedi, vici". Convertir ce message dans $\mathbb{Z}_{26}$.
2. Encoder le message avec $K = \left[\begin{smallmatrix}18 & 3\\21 & 19\end{smallmatrix}\right]$.
3. Vérifier que la matrice inverse est bien $K^{-1} = \left[\begin{smallmatrix}1 & 19\\3 & 16\end{smallmatrix}\right]$.
4. Décoder le message `YZOTEBDFKD`.

Ceci soulève plusieurs questions :

- comment trouver une clef ?
- comment inverser la clef ?

On a la proposition suivante :

Proposition 19 (Inversion d'une clef)**.** *une clef K est inversible si et seulement* $PGCD(det(K), 26) = 1$

Exemple

Prenons la clef $K = \begin{bmatrix}11 & 3\\8 & 7\end{bmatrix}$.
$\det(K) = 11 \times 7 - 8 \times 3 = 53 = 1$.
On a bien PGCD$(1, 26) = 1$. Donc, K est inversible.

Avec cette condition de validation, on peut donc :

1. tirer les composantes de la matrice K au hasard.
2. vérifier la condition PGCD(det(K), 26) = 1.
3. si elle est vérifiée, utiliser K, sinon recommencer.

Proposition 20 (Calcul de l'inverse d'une clef)**.** *si K est inversible, alors $K^{-1} = det(K)^{-1}.(K^*)^t$ ou K^* est la matrice adjointe de K (i.e. définie par $k^*_{ij} = (-1)^{i+j} det(K_{ij})$ où K_{ij} est la matrice obtenue en rayant la $i^{ème}$ ligne et la $j^{ème}$ colonne).*

Exemple

Prenons la clef $K = \begin{bmatrix} 11 & 3 \\ 8 & 7 \end{bmatrix}$.

On a déjà vérifié que cette clef était inversible.

On a $K_{11} = 7$, $K_{12} = 8$, $K_{21} = 3$, $K_{22} = 11$.

On en déduit $K^* = \begin{bmatrix} 7 & -8 \\ -3 & 11 \end{bmatrix} = \begin{bmatrix} 7 & 18 \\ 23 & 11 \end{bmatrix}$

Donc $K^{-1} = \det(K)^{-1}.K^{*t} = (1)^{-1}.\begin{bmatrix} 7 & 18 \\ 23 & 11 \end{bmatrix}^t = \begin{bmatrix} 7 & 23 \\ 18 & 11 \end{bmatrix}$

On retrouve la clef inverse déjà donnée (on avait vérifié que $K.K^{-1} = \mathrm{Id}$).

Remarque

On peut montrer (voir [Kon07], p.100) que dans $\mathbb{Z}_p$, le nombre de matrice de taille $N \times N$ inversible est :

$H_n = p^{N^2} \prod_{k=1}^{N} (1 - \frac{1}{2^k}) \simeq 0.288488.p^{N^2}$ (quand $N \to \infty$).

Donc, $N = 2$, le nombre de matrices inversibles est de 171366.

Il faut prendre des matrices de taille au moins 4 pour commencer à disposer d'une taille d'espaces de clef suffisamment dissuasive.

3.5 Permutations

Nous avons déjà vu le chiffre monoalphabétique par permutation. Celui-ci consistait à permuter les symboles de l'alphabet entre eux (i.e. changer l'alphabet par une permutation de celui-ci).

Dans le cas des chiffres par permutation, il s'agit de changer l'ordre des lettres du message en utilisant une règle de permutation de la place des symboles qui le compose.

Définition 33 (Chiffre par permutations)**.** On veut coder un message M de longueur p. Soit $\mathcal{P}_p = \mathcal{C}_p = \mathbb{Z}_{26}^p$ et $\mathcal{K}_p$ l'ensemble des permutations sur $\mathbb{Z}_p$.
Un chiffrement par permutation avec $\pi \in \mathcal{K}_p$ est défini par les règles suivantes :

- pour $M \in \mathcal{P}_p$, $e_k(M) = e_k(m_0 m_1 \dots m_{p-1}) = m_{\pi(0)} m_{\pi(1)} \dots m_{\pi(p-1)}$.

- pour $Y \in C_p$, $d_k(Y) = d_k(y_0 y_1 \dots y_{p-1}) = y_{\pi^{-1}(0)} y_{\pi^{-1}(1)} \dots y_{\pi^{-1}(p-1)}$

où π^{-1} est la transposition inverse de π (*i.e.* telle que $\pi \circ \pi^{-1} = \mathrm{Id}$).

On voit que si le message est court, la permutation est un simple jeu d'anagramme. S'il est plus long, le nombre de permutations reste limité par la longueur de la clef.

Exemple

Prenons $\pi(x) \in \mathcal{K}_p$ définit par :

$$\pi(x) = \begin{cases} \lfloor x/2 \rfloor & \text{si } x \bmod 2 = 0 \\ \lfloor x/2 \rfloor + \lfloor p/2 \rfloor + p \bmod 2 & \text{si } x \bmod 2 = 1 \end{cases}$$

Pour $p = 6$, $\pi(\{0,1,2,3,4,5\}) = \{0,3,1,4,2,5\}$ ce qui revient à accoler les deux chaînes constituées d'un caractère sur deux.

M ="MONMESSAGEACODER"

x	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$\pi(x)$	0	8	1	9	2	10	3	11	4	12	5	13	6	14	7	15
M	M		N		E		S		G		A		O		E	
		O		M		S		A		E		C		D		R
Y	M	N	E	S	G	A	O	E	O	M	S	A	E	C	D	R

Y ="MNESGAOEOMSAECDR"

Ce chiffre particulier était utilisé pendant la guerre de Sécession et connu sous le nom de "rail fence" (à deux niveaux).

Exercice 36: Rail fence

1. Code le message "Tout ce qui peut être imaginé est réel".
2. Décoder le message `LOMETNUTDVNLPNEHMESEFIEEATAESE`.

Les chiffres par permutation les plus courants sont basés sur un mot clef.

1. A partir de ce mot-clef, on construit la clef en prenant les lettres du mot clef dans l'ordre et en retirant toutes les lettres déjà rencontrées (par exemple, le mot clef "PASTEQUE" fourni 7 lettres uniques "PASTEQU").
2. On note n la longueur la clef obtenue.
3. On écrit le texte clair à chiffrer dans une table à n colonnes avec un caractère par case.

4. On permute les colonnes en utilisant celle générée par la clef en triant ses lettres par ordre alphabétique (par exemple, pour la clef "PASTEQU", après tri = "AEPQSTU", donc $\pi((0,1,2,3,4,5,6)) = (1,4,0,5,2,3,6)$).
5. On construit le chiffre comme étant la suite des colonnes les unes à la suite des autres.

L'écriture formelle de la fonction de codage et de décodage offre assez peu d'intérêt dans le cadre de cours.

Ce type de code par transposition existe dans de nombreuses variations sur la façon de construire les tables et/ou la façon de les composer (par exemple en utilisant plusieurs clefs pour effectuer des transpositions multiples).

Exemple 2

M ="MONMESSAGEADECODER".

En première ligne, la clef. On remplit la table avec une lettre du texte clair par colonne :

P	A	S	T	E	Q	U
M	O	N	M	E	S	S
A	G	E	A	D	E	C
O	D	E	R			

On lit maintenant les colonnes dans l'ordre alphabétique (*i.e.* AEPQSTU) :

O G D E D M A O S E N E E M A R S C

On obtient le chiffre : OGDEDMAOSENEEMARSC

Comment décoder ?

Le problème est que les colonnes concaténées n'ont pas toutes la même hauteur. Comment comment découper le code ?

Exemple 2 (suite)

On connaît donc le code C ="OGDEDMAOSENEEMARSC", et la clef K ="PASTEQUE". La longueur de la clef étant 7 (sans les répétitions), et celle du code 18, la chiffre est constitué de 4 colonnes de longueurs 3, et 3 colonnes de longueur 2 ($18 = 2 \times 7 + 4$, donc 4 colonnes de longueur $2 + 1$, et $7 - 4$ colonnes de longueur 2).

En réécrivant le tableau, la clef et l'ordre dans lesquelles les colonnes sont prises (voir ci-dessous), on retrouve le découpage du code dans le tableau comme : C ="OGD ED MAO SE NEE MAR SC".

Clef	P	A	S	T	E	Q	U
Ordre	3	1	5	6	2	4	7
Longueur	3	3	3	3	2	2	2
	M	O	N	M	E	S	S
	A	G	E	A	D	E	C
	O	D	E	R			

Enfin, on relis le tableau ligne par ligne "MONMESSAGEADECODER".

Exercice 37: Chiffre par permutation

On veut effectuer un chiffrage en utilisant une permutation par table avec le mot-clef SANGLIER.

1. A partir du mot-clef, déterminer comment les colonnes seront permutées.
2. Coder le message "Le doute est le commencement de la sagesse".
3. Décoder le message `ATCNEEUTOAUTCMQRRRXOFLETEDILSEO`.

4 Cryptanalyse

La **totalité** de codes abordés dans la section précédente ont été conçue à une époque où :

- l'ignorance du chiffre utilisé était sa meilleure protection.
- la cryptologie, en tant que domaine des mathématiques, n'existait pas.
- la sécurité d'un chiffre n'était pas mathématiquement prouvée.
- les ordinateurs TFlops n'existaient pas.

En conséquence, ces chiffres ont tous des fragilités.

Néanmoins, ces méthodes classiques peuvent être utilisées comme brique dans une méthode de chiffrement, à partir du moment où :

- la sécurité du code ne repose pas sur ces méthodes,
- elles n'ont pour but que de rendre plus difficile la compréhension de la VRAI méthode utilisée.

Nous donnons maintenant une typologie des attaques possibles sur un chiffre.

Il est important de bien comprendre cette typologie car elle permet de définir le niveau de sécurité d'un chiffre par sa capacité à répondre aux différents types d'attaques.

4.1 Typologie des attaques

Dans cette section, on note $C_i = E_k(M_i)$ pour désigner le chiffre C_i produit par un cryptosystème E_k de clef k à partir d'un message M_i.

Toutes les attaques ont le même but : on veut identifier le cryptosystème E et

trouver la clef k (ou alternativement un algorithme) afin de déchiffrer tout code C_p utilisant la même clef k que les données dont l'on dispose.

Les différents types d'attaques (éventuellement combinées) sont les suivantes :

- **attaque du chiffre** : on a $C_1, \ldots, C_n$.
  Trouver la clef k (ou un algorithme) permettant de trouver $M_1, \ldots, M_n$, si possible être capable de déchiffrer C_{n+1}.
  scénario : interception de messages cryptés entre Alice et Robert.
- **attaque à texte clair connu** : on a des couples (C_i, M_i)
  Trouver la clef k (ou un algorithme) afin de déchiffrer tout C_{n+1} utilisant la même clef k.
  scénario : vol de la correspondance chez Alice ou chez Robert.
- **attaque à texte clair choisi** : l'hostile a possibilité de choisir les textes M_i à chiffrer, et on lui fournit le C_i correspondant.
  Trouver la clef k (ou un algorithme) afin de déchiffrer tout C_{n+1} utilisant la même clef k.
  scénario : François demande à Alice de transmettre un message à Robert. Alice fait confiance à François, mais François travaille en fait pour Manuel (qui peut intercepter les chiffres).
- **attaque à chiffre choisi** : l'hostile a possibilité de choisir les textes C_i à déchiffrer, et on lui fournit le M_i correspondant.
  Trouver la clef k (ou un algorithme) afin de déchiffrer tout C_{n+1} utilisant la même clef k.
  scénario : François demande à Robert de déchiffrer un message pour lui. Robert fait confiance à François, mais François travaille en fait pour Manuel qui lui a donné le message à déchiffrer.

Dans tous les cas, si la méthode est suffisamment robuste :

- au minimum, le changement de la clef nécessite une nouvelle analyse.
- au mieux, tous les cas d'attaque sont :
  - ◇ soit vouées à l'échec (infaisable au sens de l'informatique théorique).
  - ◇ soit prendrons un temps plus grand que la durée de vie de l'information chiffrée.

Par durée de vie d'une information, l'on entend :

- soit le temps pendant laquelle l'information est valide
  par exemple 3 ans pour le code d'accès à une carte bancaire.
- soit le temps pendant lequel cette information est sensible
  par exemple, la liste des unités engagées sur le champ de bataille pour la bataille des prochaines semaines.

Dans tous les cas, le coût humain et financier d'une cryptanalyse doit toujours :

- paraître à l'hostile plus important que le bénéfice qu'il pourrait en retirer, afin de le dissuader même de l'envisager,
- ne lui laisser que peu d'espoir d'en tirer quelque chose.

4.2 Attaques

4.2.1 Monoalphabétique

L'espace des clefs est très petit, puisque $\#\mathcal{K} - 1 = 25$ ($k = 0$ non compté).

Supposons que nous ayons le message chiffré : $Y =$ "UMAAIOMAMKZMB"

Il suffit de tester clef après clef, jusqu'à trouver un déchiffrement qui ait du sens.

k	déchiffrement	k	déchiffrement	k	déchiffrement
1	TLZZHNLZLJYLA	10	KCQQYECQCAPCR	19	BTHHPVTHTRGTI
2	SKYYGMKYKIXKZ	11	JBPPXDBPBZOBQ	20	ASGGOUSGSQFSH
3	RJXXFLJXJHWJY	12	IAOOWCAOAYNAP	21	ZRFFNTRFRPERG
4	QIWWEKIWIGVIX	13	HZNNVBZNZXMZO	22	YQEEMSQEQODQF
5	PHVVDJHVHFUHW	14	GYMMUAYMYWLYN	23	XPDDLRPDPNCPE
6	OGUUCIGUGETGV	15	FXLLTZXLXVKXM	24	WOCCKQOCOMBOD
7	NFTTBHFTFDSFU	16	EWKKSYWKWUJWL	25	VNBBJPNBNLANC
8	MESSAGESECRET	17	DVJJRXVJVTIVK		
9	LDRRZFDRDBQDS	18	CUIIQWUIUSHUJ		

Donc, $k = 8$. En moyenne, il suffit de 13 essais avant de trouver la clef.

Conclusion : un cryptosystème doit avoir un espace de clefs tellement grand que le test systématique de toutes les clefs (=force brute) ne doit pas être une solution envisageable pour déchiffrer le message.

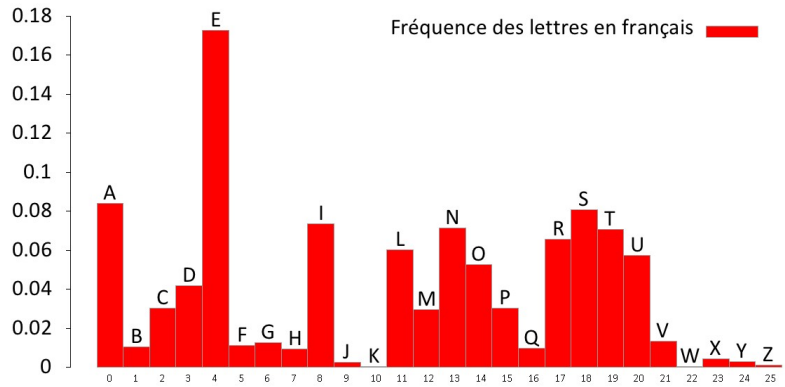
Dans un chiffre monoalphabétique, chaque lettre est remplacée par un seul symbole.

En conséquence, les caractéristiques suivantes aident à attaquer le code :

- La fréquence des lettres ne change pas
  Les symboles les plus fréquents dans le cryptogramme devraient être les mêmes que les symboles les plus fréquents dans les textes clairs.
- La fréquence des successions de symboles ne change pas
  Les symboles consécutifs fréquents (bigrammes, trigrammes) ne changent pas. Par exemple, E est suivi de S dans 19% des cas, de N dans 12%.
- Si on pense qu'un mot est présent dans le message
  Si l'hypothèse est vérifiée, on déchiffre l'ensemble des lettres du mot.
- Une fois que des portions de mots sont trouvées, on peut utiliser un dictionnaire afin de limiter le nombre de choix possible.

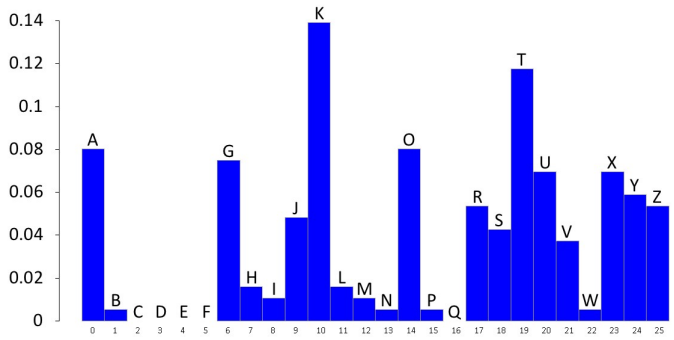
Évidemment, plus le code est long (ou plus on dispose de textes utilisant la même clef), plus les fréquences des lettres du langage du texte clair (i.e. histogramme des fréquences des lettres) suffisent pour décoder un code.

Exemple : pour le français, la fréquence des lettres est la suivante :



Utilisation sur un chiffre par décalage :

```
JKSGOTSGZOTPKTBKXXGOATKVKZOZOUTGRATJKYVKXYUTTGMKYRKYVRAYOTLRAKT
ZYJASOTOYZKXKJKRGMAKXXKGATNUSSKWAOTKVKAZXOKTXKLAYKXGRGLORRKJAHG
XUTJKVOUSHUTUAYUHZOKTJXUTYATVGXJUTZGIOZKVUAXRKIUSSGTJGTZRUAOY
```



L'histogramme pour ce message est représenté ci-contre. Le maximum est en K=10 (au lieu de E=4). On en déduit : $k = K - E = 10 - 4 = 6$.

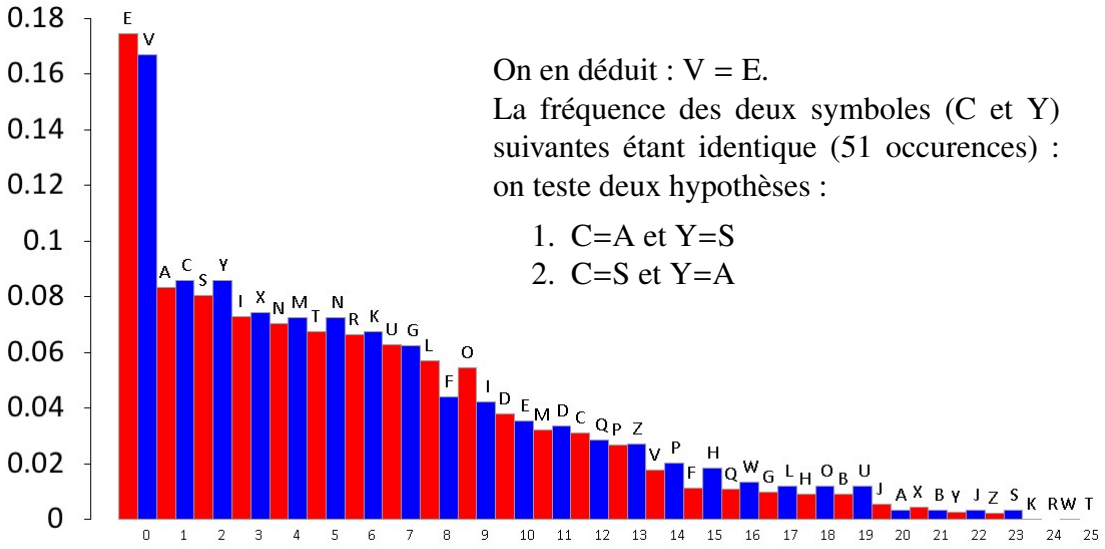
Dans le cas des chiffres par substitution, en utilisant la combinaison de l'ensemble des caractéristiques de fréquence et des transitions entre lettre, les chiffres sont

également déchiffrable sans trop de difficultés.

Considérons le code suivant :

```
EYDUVMVYEVIGXIZYWNVKYXBDGKLXVIVCLXVHYNOYNIVCHVZNVKCPVKYHGXIVMXK
VQMVDNVXCVHYNMVKPXCVMZNDVYFXKPVCQVMCGKKYJVCFVCQFXCNKOFXVKICYXBV
FVDINGKCDGEEVFVCDMNINLXVCLXNOGKIFVCMVQXIYINGKCCYKCHYEYNCQGXZGNM
CVKOYNMVXKVNFOYNIFVCPVQXIVCCYKCQGXZGNMHYEYNCFVPVZVKNMFVWMYZVUGE
EVYZGXFXEVIVEGNJKVMCYMVDGKKYNCCYKDVYWGKEYMDUVQMVCLXVCYKCWGXMCVP
VFNVMVKEVPNCYKIZGXFVSZGXCYFFVMYFYDUYEWMVHVQXNCZGXCOYNMVKGEEVMPV
QXIVCNHVEVMVCGFZYNCYVKIMVMPYKCFYDYMMNVMVQGFNINLXVFXNYNHVMVQGKPX
IMVCUAQGDMNIVEVKIDVCVMYNIQGXMEVZGXVMYFYDGEIVLXVHYNEVVIGXHVCXNCY
QQMVDNVVUWNVKKGXCZGXCPVDNPVMGKCVIKGXCYXMGKCQYMZGXCXKVNKOFXVKDVY
FYDUYEWMVDYMZGXCAWMNFFVMVS
```

On en tire l'histogramme de ses symboles suivant (en bleu) avec (en rouge) celui des lettres.



On en déduit : V = E.
La fréquence des deux symboles (C et Y) suivantes étant identique (51 occurences) : on teste deux hypothèses :

1. C=A et Y=S
2. C=S et Y=A

On écarte l'hypothèse 1, car la ligne 4 contient la suite :

`NCQGXZGNMCVKOYNMVXKVNFOYNIFVCPVQXI`**`VCCY`**`KCQGXZGNMHYEYNCFVPVZVK`

Or `VCCY` = `EAAS` est moins probable que `VCCY` = `ESSA`. On garde l'hypothèse 2.

Avec ce remplacement, on voit sur la ligne 5 :

`NMFVWMYZVUGEEVYZGXFXEVIVEGNJKVMCYMVDGKK`**`Y`**`N`**`CCY`**`KDVYWGKEYMDUVQMV`

Or `YNCCY` = `A?SSA`. Le `N` est presque sûrement un `I` et (`V`, `C`, `Y`, `N`) = (`E`, `S`, `A`, `I`).

Les suites CCY=SSA (il y en a 3) se terminent probablement par N, dont K=N.

A cet étape de la recherche, le chiffre ne contient plus aucune suite facilement

identifiable (ci-dessous en n'y laissant que les symboles identifiés pour le moment) :

```
_a__e_ea_e_____a_iena____n__e_es__e_ai_ai_es_e_iens_ena____e
__ne__e_ie_se_ai_en__se__i_ea__n_es_e_s_nna_es_es___sin___en
_sa__e_e__i_ns____e_es__i_i__es__i__n__es_e___a_i_nssans_a_a
is_____i_sen_ai_e_nei__ai__es_e___essans_____i__a_ais_e_e_en
i__e__a_e____ea______e_e__i_ne_sa_e__nnaissan_ea__n_a___e__e
s__esans____se_e_ie_en_e_isan_____e____sa__e_a_a__a___e_e__i
s___s_ai_en___e__e___esi_e_e_es___aisaen__e__ans_a_a__ie_e__
_i_i__e__iai_e_e__n____es______i_e_en__ese_ai______e___e_a_a
____e__e_ai_ee____es_isa___e_iee__ienn__s___s_e_i_e__nse_n__
sa___ns_a____s_nein___en_ea_a__a___e_a____s___i__e_e_
```

On voit avec les fréquences proches les plus élevées suivantes, que les symboles de {X, M, K, G, F, I} sans doute parmi {N, T, R, U, L, O}.

Avec les bigrammes pour les 5 symboles les plus fréquents (avant et après) avec des statistiques de plus de 10% (ou 3 symboles) sont :

	Avant	Après
Lettre E	D=14% L=14% R=11% T=10%	S=19% N=12% R=9%
Symbole V	M=18% EF=10%	C=17% M=16% K=13%
Lettre S	E=41% I=11% N=10% S=9%	E=17% A=10% D=9% S=9%
Symbole C	V=33% X=20% K=18% N=14%	Y=22% V=12% Q=10%
Lettre A	L=15% RT=11% S=10%	N=16% I=15% R=11 L=8%
Symbole Y	C=22% HVEF=10%	N=25% K=12% DEFM=10%
Lettre I	A=17% T=14% R=8%	T=17% E=14% S=12% LN=11%
Symbole N	Y=30% IMX=9% D=7%	C=16% V=16% I=14% M=14%
Lettre N	E=30% O=23% A=19% I=11%	T=23% E=14% S=12% D=11%
Symbole K	V=33% G=25% Y=15% X=10%	C=23% I=13% V=10% Y=10%

Stratégie : partir des pourcentages les plus hauts et les confirmer avec d'autres.
Rappel : (V, C, Y, N) = (E, S, A, I) et {X, M, K, G, F, I} ≃ {N, T, R, U, L, O})

Interprétation :

- Par N/K-, G=O (23%).
- Par N/K+, I = T, renforcé par I/N+ et E/V+ car M≠T.
- Par E/V-, A/Y-, A/Y+, EF=L. On prend F=L (car E moins fréquent).
- Possible : M=R (par E/V-, EV+, I/N-, A/Y+ mais pas par A/Y-, I/N+).

Vérifions les conclusions précédents : (G, I, F, M) ≃ (O, T, L, R).

```
_a__erea_eto_t_a_iena___on__etes__e_ai_aites_e_iens_ena_o_te
r_ne_re_ie_se_airen__ser_i_eal_n_es_ersonna_esles_l_sin_l_en
tsa__ele_tions_o__eles_riti__es__i_ontlesre__tationssans_a_a
is_o__oirsen_aire_neil_aitles_e__tessans_o__oir_a_aisle_e_en
irle_ra_e_o__ea_o_l__ete_oi_nersare_onnaissan_ea_on_ar__e_re
s__esans_o_rse_elieren_e_isant_o_le__o_sallerala__a__re_e__i
s_o_s_aireno__er_e__tesi_e_eresol_aisaentrer_ansla_arriere_o
liti__el_iai_ere_on__tres___o_rite_ent_eserait_o_r_e_o_erala
_o_te__e_ai_eeto__es_isa__re_iee__ienno_s_o_s_e_i_eronsetno_
sa_rons_ar_o_s_nein_l_en_eala__a__re_ar_o_s__rillere_
```

Le reste s'obtient pas reconnaissance et complétion successive :

- "reDonnaissanDe" (L5) et "Darriere" (L7) implique D=C.
- "entrerPanslacarriere" (L7) implique P=D.
- "inflXence" (L10) et "inflXents" (L1-2) implique X=U.
- "runeQrecieuseHairenduserZicealundesQersonnaJeslesQlusinfluen" (L2), implique Q=P, H=J, Z=V, J=G.
- ...

Pour chaque nouveau caractère trouvé, le nombre d'informations nouvelles est considérable (caractères à proximité, transitions, mots reconnaissables ...).

Seuls les premiers caractères posent des difficultés à trouver.

Remarques

- le style de rédaction du texte peut faire beaucoup varier les fréquences des lettres.
  Par exemple, en style télégraphique, les articles, déterminants, ...
- la loi des grands nombres assure que sur un chiffre suffisamment long (ou un corpus de textes utilisant la même clef) que l'on tendra vers la répartition exacte.
  mais sur des textes courts, les fréquences peuvent être fort différentes.
- le texte peut être volontairement écrit de manière à introduire un biais dans les fréquences.

On peut utiliser des modèles statistiques (maximum de vraisemblance) ou des modèles de Markov cachés (voir [Kon07], pp 73-90).

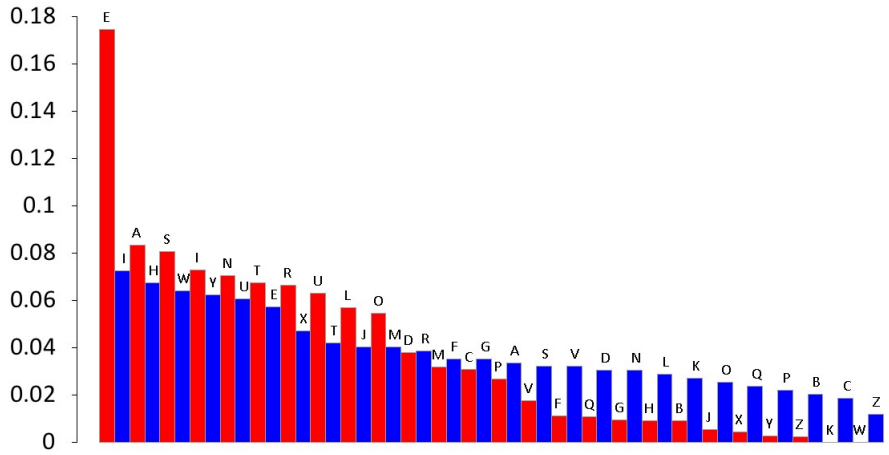
Les tests d'hypothèses effectués précédement peuvent être automatisé en partant d'abord des hypothèses les plus probables.

4.2.2 Polyalphabétique

Le chiffre de Vigenère a été considéré comme un code sûr pendant presque 3 siècles.

L'attaque basée sur la seule analyse des fréquences des symboles ne donne plus d'information.

La figure ci-dessous compare l'histogramme du texte clair précédent (en rouge) et l'histogramme du chiffre de Vigenère correspondant avec la clef "PASTEQUE".



Le cycle de décalage de la clef sur les lettres a tendance à mieux répartir les symboles sur l'ensemble des fréquences.

En 1863, Friedrich Kasiski proposa un test (qui porte maintenant son nom) permettant de déterminer la longueur de la clef.

L'idée du test de Kasiski est simple :

- la longueur p de la clef est constante et réutilisée cycliquement pour chiffrer le texte (rappel : $e_k(m_i) = m_i + k_{i \bmod p}$).
- dans un texte, les suites d'au moins trois lettres qui se répète sont rélativement fréquentes (ent, les, des, que ait, lle, ion, eme, ...).

Il arrive presque nécessairement qu'une suite quelconque d'au moins trois lettres se retrouve être codée avec exactement la même partie de la clef.

En conséquence, ces portions de texte se retrouvent à être chiffrée avec exactement les mêmes symboles, et sont identiques.

Pour la $i^{\text{ème}}$ répétition, si on note d_i la distance entre les deux motifs de symboles identiques, alors on a nécessairement : $d_i = k_i.p$ (où k_i est inconnu).

Mais, si l'on considère l'ensemble des $\{d_1, d_2, \ldots, d_k\}$ trouvés dans un texte, on

est certain que p divise tous les d_i.

En conséquence, $p = \text{PGCD}(d_1, d_2, \ldots, d_k)$

Notons qu'au pire, si l'on ne trouve pas p, on trouvera l'un de ses multiples (ce qui limitera le nombre de choix possibles pour cette longueur).

En reprenant le chiffre de Vigenère du texte précédent, on trouve 4 répétitions de motifs de taille 4 ou plus :

- `JFIH` (position 157 et 205), $d_0 = 48$
- `WIYJ` (position 162 et 210), $d_1 = 48$
- `HAFL` (position 216 et 304), $d_2 = 88$
- HVGNW (position 360 et 520), $d_3 = 160$

```
BAUAIHYEBELHYJPEQIWGEKRGDNINIJYWFUWCEYZEXTWLNUPMTNKWIDUNDULX
VKHIERWVMUOWTJSBVUHHJSWKZYWIPLMGHUMTTRKHRDUKTSDXWFFYHIFYPKYR
ISSNBUFIRTAHRIWSBMWEIIWVXTAJYUMUJIXHRJFIHRWIYJUXXOFLWQHWYAET
MIJSJVGBVIYRUAAKIKHIXLXTMJFIHDWIYJYWHAFLTEOZDIJCECUMHLWWILYR
XRDXFHUZTHGFQUUZDUDNQUNIBOAZRULWPRWVSDHEXSKTRSYEQOFFEHWLTPJX
WGOIHAFLFEOVHEVXPYYVTNEXHYMECTNHYBYDKOMLEBFIGADTGXUQQRWCIFOM
HVGNWVUMGEFHQCYVSEHNXUMMYEEXVUMSAVSBWQYRIRWKHQHWAAUTVHCIGEHH
PYNMFUWEYYUMYEJXTEHHJTJXWXSTDCJBXUGICTUXWULEXTHHYHGIKOMXVQFE
ROEMIGOIYAAFIUNSJJWLYYMEEPJXGYYIWBAXRDIYHVGNWTYGXDWKSDMIINGN
WQOVDNKIEHPSJSMGIYHJAUWGGUUPPCZTQRLIRAJOSKMCQRAEPULI
```

PGCD(48, 88, 160) = 8 est bien la longueur de notre clef (`PASTEQUE`).

A noter que pour des motifs de taille 3 (ou plus), on trouve 53 répétitions (soit environ une tous les 10 caractères !).

Lorsque l'on connait la longueur p de la clé $k = k_0 \ldots k_{p-1}$, on remarque dans le message $m = m_0 \ldots m_{n-1}$:

- $m_0, m_{0+p}, \ldots, m_{0+k.p}$ ont été chiffré par décalage avec la clef k_0.
- $m_1, m_{1+p}, \ldots, m_{1+k.p}$ ont été chiffré par décalage avec la clef k_1.
- ...
- $m_{p-1}, m_{p-1+p}, \ldots, m_{p-1+k.p}$ ont été chiffré par décalage avec la clef k_{p-1}.

Afin de trouver la clef, on utilise un outils proposé par Friedman [Fri87]. Dans texte M en langue X, soit n_i le nombre de symboles s_i, et $n = \sum_i n_i$.

Définition 34 (Index de coïncidence)**.** L'index de coïncidence I_c du texte M est la probabilité que deux symboles s_i et s_j choisis au hasard soient identiques.

$I_c(X) = \frac{n_i.(n_i-1)}{n.(n-1)} \simeq \sum_i f_i^2$

où f_i est la fréquence du symbole du s_i.

Considérons maintenant un chiffre monoalphabétique. Nous avons vu que ce chiffre ne changeait pas la répartition des caractères.

Donc si $\left\{\tilde{f}_i\right\}$ est la fréquence empirique du texte clair (*i.e.* $\tilde{f}_i = n_i/n$), il existe une permutation π telle que la fréquence du cryptogramme soit obtenue avec $\left\{\tilde{f}_{\pi(i)}\right\}$.

Etudions le comportement de $\sum \tilde{f}_{\pi(i)}\tilde{f}_i$:

Lemme 21 (Coïncidence de Friedman)**.** $\sum_i \tilde{f}_{\pi(i)}\tilde{f}_i \leq \sum_i \tilde{f}_i^2$

Démonstration:
En utilisant l'inégalité de Schwartz, on a :

$$\sum_i \tilde{f}_{\pi(i)}\tilde{f}_i \leq \sqrt{\sum_i \tilde{f}_{\pi(i)}^2}\cdot\sqrt{\sum_i \tilde{f}_i^2} \Leftrightarrow \sum_i \tilde{f}_{\pi(i)}\tilde{f}_i \leq \sum_i \tilde{f}_i^2$$

car $\sum_i \tilde{f}_{\pi(i)}^2 = \sum_i \tilde{f}_i^2$ (la permutation change juste l'ordre des termes). □

Ceci montre que l'indice de coïncidence est maximum si la permutation π est l'identité, à savoir si la permutation π est celle qui permet de déchiffrer le message.

En conséquence, les fréquences du texte clair n'étant pas connues, Friedman propose de calculer le coïncidence suivant pour évaluer une permutation π :

$$I_\pi(X) = \sum_i f_i.\tilde{f}_{\pi(i)}$$

où $\{f_i\}$ est la fréquence des lettres du langage X, $\left\{\tilde{f}_i\right\}$ la fréquence estimée des lettres sur le chiffre, et π la permutation à évaluer.

La permutation π permettant de décoder le message devrait vérifier $I_\pi \simeq I_c$.

L'indice de coïncidence I_π est donc un guide afin de limiter les tests aux permutations permettant d'obtenir les scores les plus hauts.

Dans le cas du chiffre de Vigenère, une fois la longueur p de la clef trouvée, on doit décoder p chiffres monoalphabétique par décalage, à savoir π est de la forme $\pi_k(x) = (x + k) \bmod 26$.

Tout d'abord dans le cas du français, l'indice de coïncidence est :

	A	B	C	D	E	F	G	H	I	J	K	L	M
f_i (%)	8.41	1.06	3.03	4.18	17.3	1.12	1.27	0.92	7.35	0.25	0.05	6.02	2.96
	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
f_i (%)	7.14	5.26	3.01	0.99	6.56	8.09	7.08	5.75	1.32	0.04	0.45	0.30	0.12

$I_c(X) = \sum_i f_i^2 = 0.0841 \times 0.0841 + \ldots + 0.0012 \times 0.0012 = 0.078190$

Maintenant, pour chaque lettre k_i de la clef, on va :

1. évaluer la fréquence des symboles du chiffre avec les lettres $m_{i+k.p}$ (décalé de i par pas de p),
2. calculer l'indice de coïncidence I_{π_s} pour toutes les permutations π_s avec $s = 0 \ldots 25$.

3. le s donnant la coïncidence maximale est la $i^{\text{ème}}$ lettre de la clef.

On obtient les indices de coïncidence suivants :

s	A N	B O	C P	D Q	E R	F S	G T	H U	I V	J W	K X	L Y	M Z
k_0	0.032	0.036	0.033	0.050	0.040	0.047	0.040	0.033	0.030	0.035	0.032	0.038	0.039
	0.037	0.036	0.061	0.047	0.030	0.031	0.048	0.036	0.031	0.038	0.041	0.035	0.043
k_1	0.069	0.040	0.035	0.033	0.040	0.031	0.038	0.038	0.036	0.045	0.039	0.038	0.030
	0.051	0.038	0.047	0.036	0.043	0.032	0.030	0.029	0.029	0.049	0.031	0.033	0.038
k_2	0.033	0.039	0.041	0.044	0.041	0.051	0.035	0.035	0.039	0.044	0.029	0.035	0.035
	0.029	0.045	0.032	0.035	0.035	0.081	0.040	0.034	0.030	0.055	0.026	0.026	0.030
k_3	0.035	0.030	0.039	0.046	0.040	0.043	0.042	0.041	0.034	0.043	0.039	0.035	0.029
	0.035	0.030	0.052	0.031	0.034	0.034	0.075	0.043	0.034	0.033	0.043	0.029	0.031
k_4	0.042	0.031	0.038	0.041	0.073	0.040	0.032	0.031	0.041	0.030	0.030	0.039	0.035
	0.048	0.041	0.037	0.038	0.044	0.042	0.036	0.043	0.037	0.037	0.032	0.034	0.026
k_5	0.034	0.040	0.041	0.050	0.035	0.046	0.046	0.043	0.028	0.037	0.029	0.026	0.047
	0.031	0.028	0.034	0.083	0.040	0.034	0.030	0.058	0.028	0.029	0.030	0.035	0.036
k_6	0.030	0.046	0.034	0.050	0.037	0.040	0.038	0.050	0.035	0.036	0.038	0.041	0.040
	0.036	0.035	0.022	0.051	0.024	0.039	0.032	0.080	0.037	0.035	0.027	0.037	0.030
k_7	0.057	0.030	0.032	0.032	0.083	0.036	0.031	0.028	0.054	0.027	0.031	0.030	0.031
	0.036	0.044	0.040	0.042	0.053	0.042	0.039	0.037	0.045	0.034	0.029	0.027	0.032

Les décalages s correspondant à ceux apportant les plus forts indices de correspondances sont : $(k_0, k_1, k_2, k_3, k_4, k_5, k_6, k_7)$ = (P,A,S,T,E,Q,U,E).

On retrouve bien la clef "PASTEQUE" permettant de décoder le message.

4.2.3 Polygrammiques

Ce chiffre peut être assez difficile à casser pour une attaque sur chiffre seul.

En revanche, il est assez facile d'obtenir la clef avec une attaque à texte clair connu.

Supposons que les conditions suivantes sont réunies pour lesquelles l'attaquant a :

- déterminé la taille n de bloc utilisé dans le chiffre de Hill.
- obtenu n blocs de n-gram pour lequel il dispose du chiffre y_i et de texte clair m_i associé.

On note $Y = [y_1 \dots y_n]$ et $X = [x_1 \dots x_n]$ les matrices $n \times n$ obtenues à partir des n-grams connus.

Alors, si K est la clef qui a été utilisée pour coder le chiffre, on a : $Y = K.X$.

Donc, si X est inversible, $K = Y.X^{-1}$.

Si X n'est pas inversible, il suffit de choisir un autre n-uplet de n-grams pour lequel cela est le cas.

Exemple

soit K la clef recherchée pour un chiffre de Hill de taille $n = 2$ (en modulo 26).
soient :

- $x_1 = "FR" = \left[\begin{smallmatrix} 5 \\ 17 \end{smallmatrix}\right]$ et $y_1 = "PQ" = \left[\begin{smallmatrix} 15 \\ 16 \end{smallmatrix}\right]$ (donc $y_1 = K.x_1$).
- $x_2 = "ID" = \left[\begin{smallmatrix} 8 \\ 3 \end{smallmatrix}\right]$ et $y_2 = "CF" = \left[\begin{smallmatrix} 2 \\ 5 \end{smallmatrix}\right]$ (donc $y_2 = K.x_2$).
- $x_3 = "AY" = \left[\begin{smallmatrix} 0 \\ 24 \end{smallmatrix}\right]$ et $y_3 = "KU" \left[\begin{smallmatrix} 10 \\ 20 \end{smallmatrix}\right]$ (donc $y_3 = K.x_3$).

Construisons les matrices X et Y avec les deux premières paires :
$X = \left[\begin{smallmatrix} 5 & 8 \\ 17 & 3 \end{smallmatrix}\right]$ et $Y = \left[\begin{smallmatrix} 15 & 2 \\ 16 & 5 \end{smallmatrix}\right]$

En utilisant la méthode vue dans le chiffrage de Hill, il est facile de calculer la matrice inverse : $X^{-1} = \det(X)^{-1}.(X^*)^t = \left[\begin{smallmatrix} 5 & 8 \\ 17 & 3 \end{smallmatrix}\right]^{-1} = \left[\begin{smallmatrix} 9 & 2 \\ 1 & 15 \end{smallmatrix}\right]$.

D'où l'on en déduit,
$K = Y.X^{-1} = \left[\begin{smallmatrix} 9 & 2 \\ 1 & 15 \end{smallmatrix}\right] . \left[\begin{smallmatrix} 15 & 2 \\ 16 & 5 \end{smallmatrix}\right] = \left[\begin{smallmatrix} 7 & 8 \\ 19 & 3 \end{smallmatrix}\right]$

On vérifie avec le troisième couple :
$K.x_3 = \left[\begin{smallmatrix} 7 & 8 \\ 19 & 3 \end{smallmatrix}\right] . \left[\begin{smallmatrix} 0 \\ 24 \end{smallmatrix}\right] = \left[\begin{smallmatrix} 10 \\ 20 \end{smallmatrix}\right]$

Si la taille de la clef n'est pas connue, on la teste itérativement jusqu'à obtenir un texte décodé cohérent.

Une attaque sur chiffre seul est aussi possible si l'on suppose que l'on connaît des parties du texte clair, mais sans savoir où celles-ci sont utilisées dans le chiffre (à savoir, une suite de mots de taille supérieure ou égale à $n \times n$ dont l'on pense qu'elle a été utilisée dans le message).

Il suffit alors de réutiliser la méthode précédente en testant toutes les positions possibles dans le chiffre pour la suite de mots connus (la méthode du pivot de Gauss est aussi utilisable, voir [Kon07], p102).

Conséquence : ce chiffre est très fragile dès lors que l'on réussit à obtenir un couple (texte clair,chiffre). En général, un seul message suffit pour obtenir la clef.

4.2.4 Permutations

L'attaque d'un chiffre de type "rail fence" est triviale puisqu'il s'agit juste de trouver son nombre de niveaux.

L'attaque d'une transposition basée sur des tableaux consiste :

- à prendre des portions consécutives du chiffre (que l'on pense représenter les colonnes),
- puis à les assembler en remarquant que deux colonnes consécutives devraient former un ensembles de bigrammes.
- au fur et à mesure de l'assemblage d'un nombre de colonnes consécutives correctes, des mots apparaissent, et l'on se voit conforter dans le choix de l'ordre des colonnes.

Elle soulève néanmoins les difficultés suivantes :

- on ne connaît *a priori* pas la longueur de la clef, et donc la longueur du tableau.
- le cryptogramme comportant la même fréquence de lettres qu'un texte clair (puisqu'il s'agit seulement d'une permutation), l'apparition de bigramme peut être le fruit du hasard.
- les bigrammes constitués par les coupures entre mots peuvent engendrer de bigrammes absurdes (exemple : le granD LouP Quitte).

Or, il s'avère qu'avec des moyens modernes, ces difficultés n'en sont pas vraiment :

- si les clefs sont de petite longueur, on peut tester itérativement tous les tailles n de tableau.
- on peut donner calculer des coefficients de vraisemblance des digrammes formés pour chaque couple de colonnes (il n'y en a que $n.(n-1)$).
- à partir des couples les plus vraisemblables, on peut former des chaînes de couples.

La faiblesse de ce code tient donc aux faits suivants :

- si l'on ne fait que permuter des colonnes, les caractéristiques du langage sont utilisables pour réorganiser les colonnes,
- cette méthode de construction n'utilise qu'une toute petite partie de l'ensemble des permutation possible.

5 Théorie de Shannon

Dans cette partie, nous abordons plusieurs pistes d'analyse des cryptosystèmes :

- Existe-t-il un cryptosystème parfait (à savoir qu'il est impossible de décoder) ?
- Quelle quantité d'information peut-on tirer sur la clef une fois que le chiffre est connu ?
- Les caractéristiques de la langue (i.e. de $\mathcal{P}$) a-t-elle une influence sur le cryptosystème ?
- De quel quantité de chiffres faut-il disposer pour trouver la clef ?

Ces idées ont été développées pour la première fois par Shannon en 1949 (voir

[Sha49]). La présentation de cette partie se base sur le chapitre 2 du livre de Stinson (voir [Sti06]).

5.1 Rappel de probabilité

On prend une variable aléatoire **X** définie sur un ensemble Ω et une distribution de probabilités définis sur Ω.

On a évidemment : $\sum_{x\in\Omega} \Pr[x] = 1$.

On note $\Pr[x, y]$ la probabilité que X prenne la valeur x et Y la valeur y (on parle de probabilité jointe).

Si les variables aléatoires X et Y sont indépendantes (la réalisation de l'une n'a aucun lien avec l'autre), alors :

$$\Pr[x, y] = \Pr[x].\Pr[y]$$

Sinon, $\Pr[x|y]$ est la probabilité que l'évènement x se produise, sachant que y s'est produit.

On a alors les relations suivantes :

- $\Pr[x, y] = \Pr[y|x].\Pr[x]$
- $\Pr[x|y].\Pr[y] = \Pr[y|x].\Pr[x]$ (Bayes)

Exemple

Soit un sac contenant des boules sur lesquels sont indiqués des chiffres de 1 à 9. Les chiffres pairs sont rouges, et les chiffres impairs sont verts.

Les espaces de probabilité de X (chiffre sur la boule) et de Y (couleur de la boule) sont définis respectivement sur les ensembles suivants : $\Omega_X = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$ et $\Omega_Y = \{\text{rouge}, \text{vert}\}$.

On a :

- $\Pr[x = 4] = 1/9$.
- $\Pr[y = \text{rouge}] = 4/9$.
- $\Pr[x = 4, y = \text{rouge}] = 1/9$.
- 2 tirages avec remise : $\Pr[x_1 = 4, x_2 = 4] = \Pr[x_1 = 4] \times \Pr[x_2 = 4]$. Les tirages sont indépendants.
- $\Pr[x = 4|y = \text{rouge}] = 1/4$.
- $\Pr[y = \text{rouge}|x = 4] = 1$.
- $\Pr[x = 4, y = \text{rouge}] = \Pr[y = \text{rouge}|x = 4] \times \Pr[x = 4] = 1 \times 1/9 = 1/9$.
- $\Pr[x = 4, y = \text{rouge}] = \Pr[y = \text{rouge}] \times \Pr[x = 4|y = \text{rouge}] = 4/9 \times 1/4 = 1/9$.

5.2 Loi d'un chiffre

Soit un cryptosystème $(\mathcal{P}, C, \mathcal{K}, \mathcal{E}, \mathcal{D})$.

Supposons que :

1. il existe une distribution de probabilité sur l'espace $\mathcal{P}$ des textes clairs. On note X la variable aléatoire sur $\mathcal{P}$.
2. il existe une distribution de probabilité sur l'espace $\mathcal{K}$ des clefs. On note K la variable aléatoire sur $\mathcal{K}$.

On note :
- X la loi du texte clair,
- Y la loi du chiffre.

La clef étant généralement choisie avant de connaître le message à envoyer, on suppose que X et K sont des variables aléatoires indépendantes.

Les distributions de probabilité sur $\mathcal{P}$ et $\mathcal{K}$ impliquent aussi une distribution de probabilité sur l'ensemble C des codes. On note Y la variable aléatoire sur C.

Avec cette formalisation, il est alors possible d'évaluer la probabilité $\Pr[Y = y]$ que le chiffre y soit celui transmis.

Pour une clef $k \in \mathcal{K}$, on définit l'ensemble des chiffres possible si K est la clef :
$C(K) = \{e_K(x) \mid x \in \mathcal{P}\}$

Pour un chiffre $y \in C$, on a :
$\Pr[Y = y] = \sum_{k \in K \mid y \in C(k)} \Pr[K = k].\Pr[x = d_k(y)]$
où
- la somme sur $k \in K \mid y \in C(k)$ est sur l'ensemble des clefs k qui permettent d'obtenir y comme chiffre.
- $\Pr[x = d_K(y)]$ est la probabilité pour la clef k que y soit décodé comme x.

Remarquons que pour tout $y \in C$ et $x \in \mathcal{P}$, on peut calculer la probabilité conditionnelle :
$\Pr[Y = y \mid X = x] = \sum_{k \in K \mid x = d_k(y)} \Pr[K = k]$
où la somme sur $k \in K \mid x = d_k(y)$ est sur l'ensemble des clefs k qui permettent de décoder y comme x.

Or, avec Bayes ($\Pr[x|y].\Pr[y] = \Pr[y|x].\Pr[x]$), on a :

$$\begin{aligned}\Pr[X = x|Y = y] &= \frac{\Pr[X = x].\Pr[Y = y|X = x]}{\Pr[Y = y]} \\ &= \frac{\Pr[X = x].\sum_{k \in K \mid x = d_k(y)} \Pr[K = k]}{\sum_{k \in K \mid y \in C(k)} \Pr[K = k].\Pr[x = d_K(y)]}\end{aligned}$$

qui peut être estimée dès lors que les distributions respectives sont connues.

Exemple

Soit $\mathcal{P} = \{a, b\}$ l'ensemble des messages et $C = \{1, 2, 3, 4\}$ l'ensemble des messages codés. Soit $\mathcal{K} = \{k_1, k_2, k_3\}$ l'ensemble des clefs.

On définit les distributions suivantes :
$\Pr[X = \mathcal{P}] = \left\{\frac{1}{4}, \frac{3}{4}\right\}$
$\Pr[K = \mathcal{K}] = \left\{\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right\}$

	a	b
k_1	1	2
k_2	2	3
k_3	3	4

Exemple de calcul de Pr[Y=y] :

$$\Pr[Y = 3] = \Pr[K = k_2].\Pr[x = b] + \Pr[K = k_3].\Pr[x = a]$$
$$= \frac{1}{4}.\frac{3}{4} + \frac{1}{4}.\frac{1}{4} = \frac{3}{16} + \frac{1}{16} = \frac{1}{4}$$

y	1	2	3	4
Pr[Y=y]	$\frac{1}{8}$	$\frac{7}{16}$	$\frac{1}{4}$	$\frac{3}{16}$

Exemple de calcul de Pr[X=x|Y=y] :

$$\Pr[X = b|Y = 2] = \frac{\Pr[x = b].\Pr[K = k_1]}{\Pr[K = k_1].\Pr[x = d_{k_1}(2)] + \Pr[K = k_2].\Pr[x = d_{k_2}(2)]}$$
$$= \frac{\frac{3}{4}\frac{1}{2}}{\frac{1}{2}.\frac{3}{4} + \frac{1}{4}.\frac{1}{4}} = \frac{\frac{3}{8}}{\frac{7}{16}} = \frac{6}{7}$$

valeurs de Pr[X=x | Y=y] :

	Y=1	Y=2	Y=3	Y=4
X=a	1	$\frac{1}{7}$	$\frac{1}{4}$	0
X=b	0	$\frac{6}{7}$	$\frac{3}{4}$	1

On voit que Pr[X=x | Y=y] $\neq$ Pr[X=x] sauf quand Y=3.

5.3 Cryptosystème parfait

Conséquence : en observant le chiffre, il est possible d'obtenir des informations sur le texte en clair.

Si c'est le cas, la conception de cryptosystème permet à un cryptanalyste d'obtenir un certain nombre d'informations à partir du chiffre, éventuellement suffisament pour le décoder.

On donne donc la définition suivante :

Définition 35 (Cryptosystème parfait)**.** Un cryptosystème est à secret parfait si $\forall x \in \mathcal{P}$, $\forall y \in C$, $\Pr[x|y] = \Pr[x]$.

Autrement dit,

- la connaissance du chiffre n'apporte aucune information *a priori* sur le texte clair.
- le chiffre n'est pas décodable sans la clef.

Inversement, comme $\Pr[x, y] = Pr[x|y].Pr[y] = Pr[y|x].Pr[x]$, ceci implique que $Pr[y|x] = Pr[y]$. Donc la distribution des symboles du texte clair n'a pas d'influence sur la distribution du chiffre.

Cette définition est essentiellement formelle, car en général, les distributions de X et Y sont difficilement calculables.

Théorème 22 (Chiffre à décalage parfait)**.** *Soit $\mathcal{P} = C = \mathcal{K} = \mathbb{Z}_n$. On considère le chiffre à décalage défini par la fonction de chiffrement $e_k(x) = (x + k) \bmod n$ telle que $\forall k \in \mathbb{Z}_n, \Pr[K = k] = 1/n$.*
Alors ce chiffre à décalage est à secret parfait.

Démonstration:
Soit $y \in \mathbb{Z}_n$,

$\Pr[Y = y] = \sum_{k\in\mathbb{Z}_n} \Pr[K = k].\Pr[x = d_k(y)] = \sum_{k\in\mathbb{Z}_n} \frac{1}{n}.\Pr[X = y - k]$

Or, y étant fixé, la valeur de $(y - k) \bmod n$ est une permutation de $\mathbb{Z}_n$.
Donc, $\sum_{k\in\mathbb{Z}_n} \Pr[X = y - k] = \sum_{k\in\mathbb{Z}_n} \Pr[X = x] = 1$.
En conséquence, $\Pr[Y = y] = 1/n$.

Pour tout x, y, il existe une unique clef k telle que $e_k(x) = y$ qui est $k = (y - x) \bmod n$. Donc :

$\Pr[y|x] = \Pr[K = (x - y) \bmod n] = 1/n$ (par hypothèse sur K).

En utilisant maintenant le th. de Bayes,

$\Pr[x|y] = \Pr[x].\Pr[y|x]/\Pr[y] = \Pr[x].(1/n)/(1/n) = \Pr[x]$.

Ce chiffre est donc parfait. □

Le théorème précédent peut être généralisé si :

- $\mathcal{E}$ vérifie que, $\forall x \in \mathcal{P}$, il n'existe pas deux clefs distinctes $(k_1, k_2) \in \mathcal{K}^2$ telles que $e_{k_1}(x) = e_{k_2}(x)$ (pour tout couple (x, y), il n'y a qu'une seule clef $k \in \mathcal{K}$ telle que $e_k(x) = y$).
- toutes les clefs sont équiprobables.

Le cryptosystème donné ne permettait de chiffrer des messages constitués d'un seul symbole. Ce qui conduit à la définition suivant pour des messages binaires de taille n.

Définition 36 (Masque jetable (one-time pad) ou chiffre de Vernam)**.** Soit le cryptosystème pour $n \geq 1$, et $\mathcal{P} = C = \mathcal{K} = (\mathbb{Z}_2)^n$.

Soit $x = (x_1, \ldots, x_n)$ et $k = (k_1, \ldots, k_n)$.
On définit :
$e_k(x) = (x_1 \oplus k_1, \ldots, x_n \oplus k_n)$
$d_k(y) = (y_1 \oplus k_1, \ldots, y_n \oplus k_n)$
où $\oplus$ est le ou-exclusif.

Un chiffre de Vernam est un cryptosystème parfait si toutes les clefs sont équiprobables (par application du théorème précédent sur chacun des symboles).

Le chiffre de Vernam est un chiffre à décalage pour lequel le décalage est différent pour tous les caractères de la chaîne.

Ce cryptosystème parfait a de nombreux inconvénients :

1. la clef doit être aussi longue le texte clair.
2. la clef doit être transmise par un moyen sûr.
3. la cryptosystème n'est parfait que si chaque clef n'est utilisée qu'une seule fois.

Les conséquences lorsque ce cryptosystème est utilisé sans respecter l'exigence d'utilisation unique de la clef sont que :

- le code devient fragile : une cryptanalyse sur l'ensemble des messages utilisant la même clef (voir Vigenère).
- le code ne résiste pas à une attaque à chiffre clair connu (on en déduit immédiatement la clef).

En pratique, ce cryptosystème n'est pas utilisable à grande échelle : une nouvelle clef doit être transmise pour chaque nouvelle communication, ce qui n'est pas envisageable.

On souhaiterait pouvoir chiffrer avec une clef un chaîne relativement longue tout en conservant un niveau de sécurité suffisant.

5.4 Rappel de théorie de l'information

On rappelle les notions déjà abordées en théorie de l'information.

Entropie : $H(X) = -\sum_{x \in X} \Pr[x] \log_2 \Pr[x]$

L'entropie est plus une mesure de l'incertitude sur X plus que de l'information contenue dans X (voir l'exercice sur l'information négative).

Entropie jointe : $H(X, Y) = -\sum_{y \in Y} \sum_{x \in X} \Pr[x, y] \log_2 \Pr[x, y]$
Propriété : $H(X, Y) \leq H(X) + H(Y)$

avec égalité si les v.a. sont indépendantes.

Entropie conditionnelle :

$H(X|y) = -\sum_{x\in X} \Pr[x|y] \log_2 \Pr[x|y]$

$H(X|Y) = -\sum_{y\in Y} \sum_{x\in X} \Pr[y] \Pr[x|y] \log_2 \Pr[x|y]$

Propriétés :

$H(X, Y) = H(Y) + H(X|Y) = H(X) + H(Y|X)$.

$H(X|Y) \leq H(X)$ avec égalité si les v.a. X et Y sont indépendantes.

Comment lier les informations issues des différentes composantes d'un cryptosystème ?

Avant, nous devons d'abord voir comment mesurer la quantité d'information partagée entre deux v.a.

Définition 37 (Information d'un évènement)**.** L'information apportée par l'évènement E sur une v.a. X est définie par :

$i(E|X) = H(X) - H(X_E)$.

Considérons alors deux cas :

- Si E correspond à l'observation d'une valeur x de X (*i.e.* $E = x$), alors $H(X_E) = H(X|x) = 0$, d'où $i(x|X) = H(X)$. Autrement dit, X n'apporte aucune information sur lui-même.
- Pour un évènement E quelconque, on a :
  $H(X_E) = -\sum_{x\in X} \Pr[X|E] \log_2 \Pr[X|E]$.

Exemple

On tire un dé. Quelle est l'information apportée par E = le résultat est pair ?

$H(X) = -6 \times \frac{1}{6} \log_2(\frac{1}{6}) = \log_2(6) = 2.58$ bits

$\Pr[X|E] = \frac{1}{3}$ si $X \in E$ et 0 sinon, donc :

$H(X_E) = -\sum_{x\in\{1,3,5\}} \Pr[x|E] \log_2 \Pr[x|E] - \sum_{x\in\{2,4,6\}} \Pr[x|E] \log_2 \Pr[x|E]$

$= -3 \times 0 \log_2(0) - 3 \times \frac{1}{3} \log_2(\frac{1}{3}) = \log_2(3) = 1.58$ bit

Donc, $i(E|X) = H(X) - H(X_E) = \log_2(6) - \log_2(3) = \log_2(2) = 1$ bit.

Mais quelle est l'espérance de l'information lorsque nous apprenons si E se réalise (ou pas) ?

Autrement dit, en moyenne, quelle est l'information apportée par E ?

Il y a alors deux possibilités : il a lieu ou non. On crée alors une nouvelle v.a. Y définie par : $p_Y(E) = p_X(E) = \sum_{x\in E} p_X(x)$ et $p_Y(\overline{E}) = p_X(\overline{E}) = \sum_{x\in\overline{E}} p_X(x)$.

On calcule alors son espérance :

$$\begin{aligned} E_Y(i(Y|X)) &= p_Y(E).i(E|X) + p_Y(\overline{E}).i(\overline{E}|X) \\ &= p_Y(E).\left(H(X) - \sum_{x\in X} \Pr[x|E] \log_2 \Pr[x|E]\right) \\ &\quad + p_Y(\overline{E}).\left(H(X) - \sum_{x\in X} \Pr[x|\overline{E}] \log_2 \Pr[x|\overline{E}]\right) \\ &= H(X) - \sum_{y\in Y}\sum_{x\in X} \Pr[y].\Pr[x|y] \log_2 \Pr[x|y] \\ &= H(X) - H(X|Y) \end{aligned}$$

Exemple : en reprenant l'exemple précédent :

On a : $\Pr[E] = \Pr[\overline{E}] = 1/2$ et $H(X_{\overline{E}}) = log_2(3)$.
Donc $E_Y(i(Y|X)) = \log_2(6) - (\frac{1}{2}\log_2(3) + \frac{1}{2}\log_2(3)) = \log_2(2) = 1$ bit.

Définition 38 (Information mutuelle)**.** On appelle l'information mutuelle $I(X|Y) = H(X) - H(X|Y)$

L'information mutuelle est la mesure de l'information moyenne gagnée sur X en observant Y. Cette information est toujours positive (découle de $H(X|Y) \leq H(X)$).

Remarque 2 : sur l'information mutuelle

L'information mutuelle vérifie les propriétés suivantes :

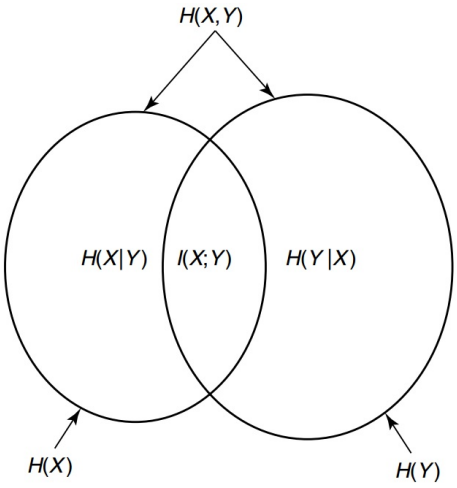
- $I(X|Y) \geq 0$ (égalité si X et Y sont indépendants).
- $I(X|Y) = H(X) + H(Y) - H(X,Y)$ (découle de $H(X,Y) = H(Y) + H(X|Y)$).
- $I(X|Y) = I(Y|X)$ (découle de la relation précédente).
  On peut donc noter $I(X,Y)$ en lieu et place de $I(X|Y)$ et $I(Y|X)$.

L'information mutuelle est donc bien mutuelle (la quantité d'information apportée

par X sur Y et égale à la quantité d'information apportée par Y sur X).

Les relations entre :
- l'entropie
- l'entropie conditionnelle
- l'entropie jointe
- l'information mutuelle

de deux v.a. X et Y peuvent être représentées sur une figure (voir ci-contre).



On note maintenant :
- K la loi de la clef,
- P la loi du texte clair,
- C la loi du chiffre.

Que peut-on dire pour un cryptosystème ?
- les fonctions de chiffrement et de déchiffrement sont déterministes. Donc : $H(C|P,K) = 0$ et $H(P|C,K) = 0$.
- pour un couple (texte clair,chiffre) donné, il n'y a qu'une seule clef. Donc, $H(K|P,C) = 0$.
- la clef est indépendante du chiffre, donc $H(K,P) = H(K) + H(P)$.
- si un cryptosystème est parfait, $\mathcal{P}$ et C sont indépendants et donc $I(P,C) = 0$.
- Par définition, on a $H(C,K) = H(K) + H(C|K)$. Donc, en considérant P en plus, on obtient : $H(C,K,P) = H(K,P) + H(C|K,P)$.

5.5 Equivocité d'une clef

Théorème 23 (Equivocité de la clef d'un cryptosystème $(\mathcal{P}, C, \mathcal{K}, \mathcal{E}, \mathcal{D})$)**.** $H(K|C) = H(K) + H(P) - H(C)$

L'équivocité d'une clef est la quantité d'incertitude restant sur la clef une fois que le cryptogramme est connu.

Démonstration:

Tout d'abord, remarquons que $H(K,P,C) = H(C|K,P) + H(K,P)$.

Comme $H(C|P,K) = 0$ (voir ci-dessus), donc $H(K,P,C) = H(K,P)$.

Mais K et P sont indépendants, donc $H(K,P) = H(K) + H(P)$.

Ainsi, $H(K,P,C) = H(K,P) = H(K) + H(P)$.

De manière symétrique, $H(K,P,C) = H(P|K,C) + H(K,C)$.
Comme $H(P|K,C) = 0$ et que K et C sont indépendants, $H(K,C) = H(K) + H(C)$.
Ainsi, $H(K,P,C) = H(K,C) = H(K) + H(C)$.

Maintenant, on a : $H(K|C) = H(K,C) - H(C) = H(K,P,C) - H(C)$.
D'où l'on conclut : $H(K|C) = H(K) + H(P) - H(C)$. □

Comment la cryptanalyse peut-elle utiliser ce résultat ?

5.6 Entropie et redondance du langage naturel

L'entropie de la clef connaissant le chiffre dépend fortement de l'entropie du texte clair. Que peut-on dire de ce dernier ?

Ici, P représente une variable aléatoire dont la réalisation produit des textes en langage naturel L.

Définition 39 (Entropie d'un langage naturel L)**.** L'entropie de L est définie comme :
$H_L = \lim_{n\to\infty} \frac{H(P^n)}{n}$

Définition 40 (Redondance d'un langage naturel L)**.** La redondance de L est définie comme :
$R_L = 1 - \frac{H_L}{log_2|\mathcal{P}|}$

Les résultats empiriques ont donnés une valeur de $H_L \simeq 1.25$.
Ce qui signifie que la redondance du langage est d'environ 75%.

Quel est l'impact sur un cryptosystème pour un chiffre de longueur n ?

5.7 Nombre de clefs erronnées

Si on obtient avec un chiffre monoalphabétique le crypotogramme le mot `WCCWC`, alors en langage naturel, il n'y a que deux solutions possibles : `ERRER` et `ESSES`.

Dans ce cas, le langage naturel sous-jacent à P a donc réduit considérablement l'espace des clefs possibles à quelques unes.

Mais que se passe-t-il dans le cas général ?

Pour tout chiffre $y \in C^n$, on définit :
$K(y) = \{k \in \mathcal{K} \mid \exists x \in \mathcal{P}^n \text{ où } \Pr[x] > 0 \text{ et } e_k(x) = y\}$.

qui est l'ensemble des clefs k dont pour lequel x est un texte du langage, et le chiffrement de x par k donne y.

Évidemment, un seul élément de $K(y)$ est la bonne clef.

Calculons maintenant la taille moyenne des ensembles de ces ensembles (réduit de 1 élément = la bonne clef).

Le nombre moyen $\overline{s}_n$ de clefs éronées est :
$\overline{s}_n = \sum_{y \in C^n} \Pr[y].(|K(y)| - 1)$

Essayons de borner supérieurement le nombre moyen de clefs erronées afin de savoir comment l'augmentation de n agit sur cet ensemble.

Remarquons que :

$$\begin{aligned} H(K|C^n) &= \sum_{y \in C^n} \Pr[y].H(K|y) \\ &\leq \sum_{y \in \mathbf{C}^n} \Pr[y]. \log_2 |K(y)| \text{ (majoration par max. entropie)} \\ &\leq \log_2 \left(\sum_{y \in \mathbf{C}^n} \Pr[y].|K(y)| \right) \text{ (par l'inégalité de Jensen)} \\ &\leq \log_2(\overline{s}_n + 1) \text{ par définition de } \overline{s}_n \end{aligned}$$

Rappel (inégalité de Jensen), si f convexe et $\sum_i p_i = 1$, $\sum_i p_i f(x_i) \geq f(\sum_i p_i.x_i)$.

Cherchons une minoration de $H(K|C^n)$,

Notons tout d'abord que $H(K|C^n) = H(K) + H(P^n) - H(C^n)$.

Remarquons que pour n assez grand :
$H(P^n) \simeq n.H_L = n.(1 - R_L). \log_2 |\mathcal{P}|$

Par ailleurs, $H(C) \leq \log_2 |C|$ (symboles du chiffre équiprobables)
Donc, $H(C^n) \leq n. \log_2 |C|$.

Si $|C| = |\mathcal{P}|$ (noté ℓ), on en déduit :

$$\begin{aligned} H(P^n) - H(C^n) &\geq n.(1 - R_L). \log_2(\ell) - n \log_2(\ell) \\ &\geq -n.R_L. \log_2(\ell) \end{aligned}$$

D'où $H(K|C^n) \geq H(K) - n.R_L. \log_2(\ell)$.

On peut borner inférieurement $\overline{s}_n$ par :

$$H(K) - n.R_L. \log_2(\ell) \leq \log_2(\overline{s}_n + 1)$$

On a alors le résultat suivant :

Théorème 24. *Soit un cryptosystème* $(\mathcal{P}, C, \mathcal{K}, \mathcal{E}, \mathcal{D})$ *tel que* $|\mathcal{P}| = |C|$ *et tel que les clefs soient choisies équiprobables.*
Alors, pour un cryptogramme de longueur n assez grand, le nombre moyen de clefs erronées $\overline{s}_n$ *vérifie :*

$$\overline{s}_n \geq \frac{|\mathcal{K}|}{|\mathcal{P}|^{n.R_L} - 1}$$

où R_L *est la redondance du langage sous-jacent.*

Démonstration:
Utiliser la borne trouvée pour $\overline{s}_n$ avec $H(K) = \log_2 |\mathcal{K}|$. □

Quand n croît, $\frac{|\mathcal{K}|}{|\mathcal{P}|^{n.R_L}-1}$ tend exponentiellement rapidement vers 0.

Autrement dit, au fur et à mesure que n grandi, l'ensemble des clefs possibles se réduit exponentiellement rapidement vers 0.

5.8 Distance d'unicité

Définition 41 (Distance d'unicité)**.** La distance d'unicité d'un cryptosystème est la valeur n_0 de n à partir de laquelle le nombre espéré de clefs erronées devient 0.

Donc, c'est la longueur moyenne du cryptogramme nécessaire à l'attaquant pour calculer la clef (sous l'hypothèse qu'il dispose des ressources nécessaires).

Cette valeur de n peut être estimée en résolvant $\overline{s}_n = |\mathcal{K}|/(|\mathcal{P}|^{n.R_L} - 1) = 0$. On obtient :

$$n_0 \simeq \frac{\log_2 |\mathcal{K}|}{R_L . \log_2 |\mathcal{P}|}$$

Pour un cryptosystème monoalphabétique par permutation, $|\mathcal{P}| = 26$ et $|\mathcal{K}| = 26!$. Avec $R_L = 0.75$, on a :

$$n_0 \simeq \frac{\log_2(26!)}{R_L . \log_2(26)} = \frac{88,4}{0,75 \times 4,7} \simeq 25$$

Une cryptogramme de taille 25 en moyenne permet habituellement de trouver la clef.

5.9 Conclusion

Remarques :

- Ceci montre à quel point la distribution du $\mathcal{P}^n$ (et plus généralement la redondance du langage) est l'un des points faibles importants de tout cryptosystème.
- Il s'agit certe d'une borne inférieure, mais elle correspond au cas le pire (loi de Murphy !), et est réalisée dans le cas d'une permutation.
- Un chiffre de Vernam n'entre pas dans le cadre précédent (puisque la longueur de la clef augmente avec le code).

En conclusion, il faut absolument avoir conscience que :

- toute redondance est particulièrement redoutable en cryptographie : elles peuvent être immédiatement exploitées pour extraire des informations à partir d'un cryptogramme.
- il faut éviter de crypter des textes "clairs" dont les distributions de probabilité sont biaisées. Il est donc préférable d'effectuer un codage entropique des données AVANT le chiffrage afin de maximiser l'entropie (donc l'incertitude) du texte qui sera crypté.

6 Principes généraux

6.1 Cryptosystème incassable

Les premiers principes de cryptographie ont été énoncés à la fin du XIX$^{\text{ème}}$ par Auguste Kerckhoffs, auteur d'un essai sur la cryptographie militaire faisant référence.

Ils restent grandement toujours valables.

Avant d'aborder ces principes, nous devons tout d'abord définir une mesure de la fiabilité d'un système cryptographique.

Un cryptosystème est dit :

- **incassable** s'il existe aucune technique pour déterminer la clef k ou le texte clair à partir du chiffre.
  Ceci doit être mathématiquement prouvé : le code est théoriquement incassable.
- **incassable en pratique** si les efforts nécessaires pour casser le cryptosystème (en temps ou en mémoire) sont plus important que les gains que l'ennemi peut en tirer.
  En pratique, la majorité des cryptosystèmes est cassable.
  Comment affiner cette notion ?

Cryptosystème incassable en pratique

Il faut considérer qu'un code peut être cassé si l'ennemi s'en donne les moyens.

Si un code n'a pas de défaut exploitable, la sécurité d'un code dépend typiquement de la longueur n de sa clef.

- Plus la clef est longue, et plus son usage consomme du temps au chiffrement et au déchiffrement.
  Ce coût est répété à grande échelle.
- Le temps de déchiffrement est proportionnel à $t(n)$ (où $t(n)$ est une fonction estimant le temps de déchiffrement en fonction de la longueur n de la clef).
  proportionnel = suivant l'importance des moyens mis-en-oeuvre.

La question n'est donc pas de savoir si un chiffre sera cassé, mais quand il le sera, si l'ennemi s'en donne les moyens.

6.2 Cryposystème robuste

Cryposystème robuste :

- Pour un cryptosystème robuste, dire qu'un chiffre peut être cassé, signifie que l'on pourra obtenir le texte clair d'un seul message (ou au pire de tous les messages chiffrés avec cette clef).
- S'il n'est pas robuste, ce sont tous les messages qui pourront être déchiffrés.

La longueur de la clef pour un message doit donc être choisie,

- non pas pour que les informations qu'il contient ne puisse être décodées (car elles le seront si l'ennemi le décide),
- mais afin que, lorsque le temps $t(n)$ sera écoulé, les informations découvertes par l'ennemi ne soient plus valides ou sensibles (donc exploitables).

En conséquence,

- la force du cryptosystème utilisé dépend de la sensibilité des données cryptées et de leurs durées de vie.
- une donnée sensible sur le long terme ne doit donc jamais être mise en situation d'être interceptée (donc transmise par un moyen de communication où son interception est possible).

6.3 Principes de Kerckhoffs

Selon Kerckhoffs, les 6 principes qu'un cryptosystème sûr doit vérifier sont les suivants :

1. Le cryptosystème doit être, s'il n'est pas théoriquement incassable, incassable en pratique.

A savoir, les informations obtenues au bout du temps nécessaire à casser un chiffre doivent avoir, au moment où le chiffre est cassé, un intérêt nul ou négligeable.

2. Le cryptosystème ne doit pas exiger le secret. Il peut tomber au nom de l'ennemi sans inconvénient tomber entre les mains de l'ennemi.
   Ni l'information, ni la connaissance du cryptosystème ne doivent le mettre en péril. Seule la clef doit permettre de décoder un chiffre.
3. La méthode permettant de choisir la clef du cryptosystème doit être facile à mémoriser et à changer.
   Idéalement, la clef devrait être aléatoire. Pour une clef devant être mémorisée (exemple : mots de passe), prendre un compromis entre la difficulté à la mémoriser et la facilité à la deviner.

4. le chiffre doit pouvoir être transmis par télégraphe.
   aujourd'hui = à travers internet.
5. l'outil mettant en oeuvre le cryptosystème doit être portable.
   aujourd'hui = implémenté sur des microprocesseurs (portables, systèmes embarqués, ...).
6. l'utilisation ne doit pas requérir la connaissance d'une longue liste de règles ou exiger un stress mental.
   reste l'une des exigences actuelles :
   - les systèmes informatiques étant complexes, l'utilisation de réseaux sécurisés nécessitent des règles strictes et des comportements adaptés.
   - sur des réseaux de communication, la vitesse requise de chiffrement/déchiffrement peu limiter la sécurité de la transmission.

7 Conclusion

Fin de la première partie du cours sur la cryptographie.

Cette première partie du cours avait pour but de vous faire comprendre que :

- la cryptographie est une affaire sérieuse.
  elle est désormais une branche des mathématiques.
- tous les chiffres classiques doivent être considérés comme non sûr.
  sauf pour cacher des données à votre petit frère.
- il ne faut pas croire que protéger la méthode de chiffrage protège le chiffre.
  (voir Kerckhoffs) compter sur l'ignorance d'autrui est une grave erreur.

- il ne faut pas inventer ses propres chiffres.
  sauf à être mathématicien et en mesure de prouver sa robustesse.
- il ne faut jamais croire qu'un cryptosystème inviolable.
  casser un code est seulement une histoire de temps et de moyen.

Nous aborderons dans les prochains chapitres :

- le chiffrement à clef privée à travers l'exemple de DES.
- un rappel d'arithmétique entière.
- le chiffrement à clef publique à travers l'exemple de RSA.