Chap. V : Cryptanalyse du chiffrement par substitution

Laurent Poinsot

8 octobre 2009

Plan

Principe de Kerckhoff

Lorsque l'on s'intéresse aux cryptanalyses, l'hypothèse généralement faite est que l'opposant, Oscar, connaît le système cryptographique employé. Ceci est appelé le **principe de Kerckhoff**. Bien sûr, si Oscar ne connaît pas le procédé employé, sa tâche sera bien plus difficile, mais on ne souhaite pas baser la sécurité du système sur la protection (sans doute incertaine) de la description des fonctions cryptographiques. Le but est donc d'étudier les fonctions cryptographiques suivant le principe de Kerckhoff.

Objectif de la cryptanalyse

Une cryptanalyse a généralement pour but de déterminer la clef secrète utilisée lors d'un échange de messages chiffrés. Oscar ne se contente donc pas de retrouver les messages clairs, il est beaucoup plus ambitieux que cela : il doit retrouver la clef. De fait, il pourra alors retrouver tous les messages clairs qui ont été chiffrés avec cette clef !

Dans ce chapitre, on s'intéresse à une technique de cryptanalyse permettant de casser un procédé de chiffrement par substitution. Cette technique est basée sur l'analyse des fréquences d'occurrence des lettres dans un texte écrit dans une langue donnée (par exemple, l'anglais ou le français).

Dans le cas présent, on effectue une hypothèse simplificatrice : on suppose que le texte clair est un message rédigé en anglais sans ponctuations ni espaces. Plusieurs personnes ont estimé la probabilité d'apparition des 26 lettres de l'alphabet en faisant des statistiques sur de nombreux romans, magazines et journaux quotidiens. Les estimations suivantes sur la langue anglaise ont été obtenues par Beker et Piper.

Fréquences d'occurrences des lettres dans les textes écrits en anglais (Beker & Piper)

lettre	proba	lettre	proba
a	0,082	*n*	0,067
b	0,015	*o*	0,075
c	0,028	*p*	0,019
d	0,043	*q*	0,001
e	0,127	*r*	0,060
f	0,022	*s*	0,063
g	0,020	*t*	0,091
h	0,061	*u*	0,028
i	0,070	*v*	0,010
j	0,002	*w*	0,023
k	0,008	*x*	0,001
l	0,040	*y*	0,020
m	0,024	*z*	0,001

À partir de ces résultats, Beker et Piper ont classé les 26 lettres en cinq groupes :

1. ”e”, ayant pour probabilité d’enviro 0,120 ;
2. ”t”, ”a”, ”o”, ”i”, ”n”, ”s”, ”h” et ”r”, ayant une probabilité entre 0,06 et 0,09 ;
3. ”d” et ”l” ayant une probabilité d’environ 0,04 ;
4. ”c”, ”u”, ”m”, ”w”, ”f”, ”g”, ”y”, ”p” et ”b” ayant une probabilité entre 0,015 et 0,028 ;
5. ”v”, ”k”, ”j”, ”x”, ”q” et ”z” ayant une probabilité inférieure à 0,01.

Il peut être utile également d'étudier la probabilité de deux ou trois lettres consécutives, appelés **digrammes** ou **trigrammes**. En anglais, les 30 digrammes les plus fréquents sont (par ordre décroissant) ”th”, ”he”, ”in”, ”er”, ”an”, ”re”, ”ed”, ”on”, ”es”, ”st”, ”en”, ”at”, ”to”, ”nt”, ”ha”, ”nd”, ”ou”, ”ea”, ”ng”, ”as”, ”or”, ”ti”, ”is”, ”et”, ”it”, ”ar”, ”te”, ”se”, ”hi” et ”of”. Les douze trigrammes les plus fréquents sont (par ordre décroissant) ”the”, ”ing”, ”and”, ”her”, ”ere”, ”ent”, ”tha”, ”nth”, ”was”, ”eth”, ”for” et ”dth”.
Maintenant que l'on sait que les lettres n'apparaissent pas toutes avec la même fréquence, on veut tirer profit de ce biais statistique afin de cryptanalyser le procédé de chiffrement pas substitution.

On considère le texte chiffré suivant obtenu par substitution.

yifqfmzrwqfyvecfmdzpcvmrzwnmdzvejbtxcddumj
ndifefmdzcdmqzkceyfcjmyrncwjcszrexchzunmxz
nzucdrjxyysmrtmeyifzwdyvzvyfzumrzcrwnzdzjj
xzwgchsmrnmdhncmfqchzjmxjzwiejyucfwdjnzdir

L'analyse des fréquences d'apparition des lettres dans **ce** texte chiffré est donnée ci-dessous.

lettre	nbr d'app	lettre	nbr d'app
a	0	*n*	9
b	1	*o*	0
c	15	*p*	1
d	13	*q*	4
e	7	*r*	10
f	11	*s*	3
g	1	*t*	2
h	4	*u*	5
i	5	*v*	5
j	11	*w*	8
k	1	*x*	6
l	0	*y*	10
m	16	*z*	20

Comme la lettre ”z” apparaît significativement plus souvent que toutes les autres lettres, on peut conjecturer que $D_K(z) = e$. Les autres caractères qui apparaissent au moins 10 fois chacun sont ”c”, ”d”, ”f”, ”j”, ”m”, ”r” et ”y”. On peut donc supposer que ces lettres sont chiffrées à partir des caractères parmi ”t”, ”a”, ”o”, ”i”, ”n”, ”s”, ”h” et ”r”, mais leur fréquence ne varie pas suffisamment pour que l'on puisse établir une correspondance plus probable.
On peut étudier les digrammes, surtout ceux de la forme ”z-” ou ”-z”, puisque l'on a déjà reconnu la lettre ”z”. On trouve que les digrammes de ce type les plus fréquents sont ”dz” et ”zw” (4 occurrences chacun), et ”rz”, ”hz”, ”xz”, ”fz”, ”zr”, ”zv”, ”zc”, ”zd” et ”zj” (deux occurrences chacun). Comme ”zw” apparaît quatre fois et ”wz” aucune, et que ”w” apparaît moins que plusieurs autres lettres, on devine que $D_K(w) = d$.

Comme ”dz” apparaît quatre fois et ”zd” deux, on imagine que $D_K(d) \in \{r, s, t\}$, mais on ne peut pas affirmer laquelle des trois possibilités est la bonne.
Sous l'hypothèse $D_K(z) = e$ et $D_K(w) = d$, en regardant une nouvelle fois le texte chiffré, on constate que ”zrw” et ”rzw” apparaissent au début du texte, et que ”rw” apparaît plus loin. Comme ”r” apparaît souvent et que ”nd” est un digramme fréquent, on peut supposer que $D_K(r) = n$.
À ce stade du raisonnement on obtient

yifqfmendqfyvecfmdepcvmnednmdevejbtxcddumj
ndifefmdecdmqekceyfcjmynncdjcsenexcheunmxe
neucdnjxyysmntmeyifeddyvevyfeumnecndnedejj
xedgchsmnnmdhncmfqchejmxjediejyucfddjnedin

On peut ensuite essayer $D_K(n) = h$, car ”nz” est un digramme fréquent mais pas ”zn”. Si ceci est correct, le segment du texte clair ”ne-ndhe” suggère de prendre $D_K(c) = a$. Avec ces hypothèses ont obtient :

yifqfmendqfyveafmdepavmnedhmdevejbtxaddumj
hdifefmdeadmqekaeyfajmynhadjasenexaheuhmxe
heuadnjxyysmntmeyifeddyvevyfeumneandhedejj
xedgahsmnhmdhamfqahejmxjediejyuafddjhedin

On considère maintenant ”m”, le second caractère le plus fréquent. Le segment ”rnm” que l'on suppose déchiffrer en ”nh-” suggère que ”h-” est le début d'un mot. Donc ”m” est certainement une voyelle. Comme on a déjà trouvé ”e” et ”a”, on peut supposer $D_K(m) = i$ ou o. Comme ”ai” est bien plus fréquent que ”ao”, le digramme du texte chiffré ”cm” conseille de supposer $D_K(m) = i$ en premier. On a donc :

yifqfiendqfyveafidepavinedhidevejbtxaddu ij
hdifefideadiqekaeyfajiynhadjasenexaheuhixe
heuadnjxyysintieyifeddyvevyfeuineandhedejj
xedgahsinhidhaimfqahejixjediejyuafddjhedin

On cherche ensuite quelle lettre se déchiffre en ”o”. Comme ”o” est une lettre fréquente, on devine qu'elle correspond à ”d”, ”f”, ”j” ou ”y”. ”y” semble être l'hypothèse la plus probable car sinon, on obtiendrait de longues suites de voyelles comme ”aoi” à partir de ”cfm” ou ”cjm”. Supposons donc que $D_K(y) = o$.
Les trois lettres les plus fréquentes restantes sont ”d”, ”f” et ”j”, qui se déchiffrent certainement en ”r”, ”s” et ”t” dans un ordre à déterminer. Deux occurrences du trigramme ”nmd” suggèrent que $D_K(d) = s$ en donnant le trigramme ”his” (c'est d'ailleurs en accord avec l'hypothèse précédente $D_K(d) \in \{r, s, t\}$). Le segment ”hncmf” pourrait se déchiffrer en ”chair”, ce qui donnerait $D_K(f) = r$ et $D_K(h) = c$, et donc on aurait $D_K(j) = t$ par élimination.

On obtient maintenant :

oirqriendqrovearisepavinedhisevetbtxassuit
hsireriseasiqekaeorationhadtasenexaceuhixe
heuasntxoosintieoiredsovevoreuineandhesett
xedgacsinhischairqacetixtedietouardsthein

Il est maintenant très facile de retrouver le texte clair et la clef utilisée :
Our friend from Paris examined his empty glass with surprise, as if evaporation had taken place while he wasn't looking. I poured some more wine and he settled back in his chair, face tilted up towards the sun.
Ce qui signifie :
Notre ami de Paris contempla son verre vide avec étonnement, comme si son contenu s'était évaporé pendant qu'il n'y prétait pas attention. Je lui versais à nouveau du vin et il se détendit à nouveau dans chaise, le visage tourné vers le soleil.

Conclusion

Il faut trouver un moyen afin d'éviter ce type de cryptanalyse basées sur l'analyse fréquentielle des lettres. Il faut donc être capable de "masquer" ces fréquences. La solution consiste à employer des **systèmes de chiffrement par blocs itérés**.