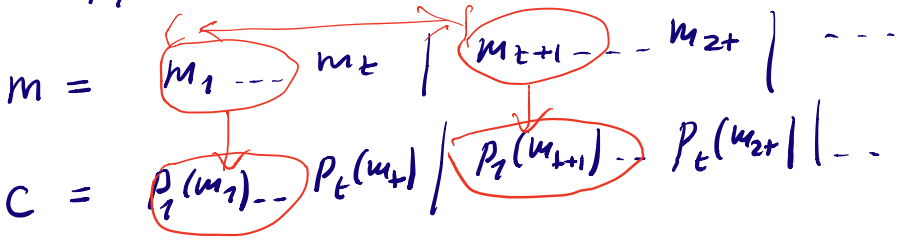
Cryptanalysis of Vigenère cipher

A alphabet, Latin

$p_1 p_2 \dots p_t$ on A, t blocksize (period)



$p_1 \dots p_t$ are shifts on A simple Vigenère

random full Vigenère

Problem. given cipher-text recover pl.-text.

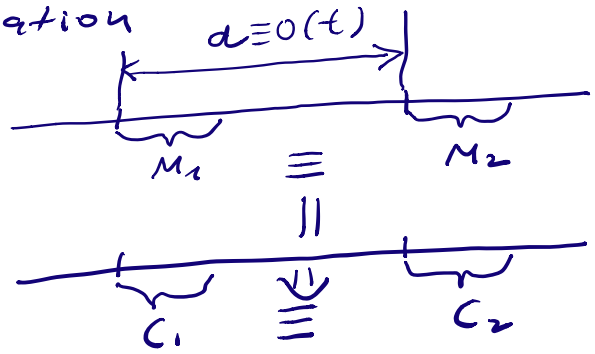
we do not know t.

1) find t
2) use frequency analysis to find m.

How to find t?

1. Kasiski test.

observation $d \equiv 0 (t)$ $\Leftarrow$ $t \mid d$ $\Leftarrow$ d is a multiple of t



search c.-text for pairs of identical segments and note distance between starting points of them

collect distances $d_1, \dots, d_s$

very likely $t = \gcd(d_1, \dots, d_s)$

some distance $d_{s+1} \not\equiv 0 (t)$

$\gcd(d_1, \dots, d_s, d_{s+1})$ very small like $1, 2, \dots$

2. Autocorrelation (more Randy)

$C = C_1 C_2 \dots C_N$

$C_i = C_{i+d}$ — # occurrences

distance d	$\frac{\#}{N-d}$	
1	≈ 0.038	$\frac{1}{26}$
2	≈ 0.038	⋮
⋯		
$t-1$	≈ 0.038	
t	first relatively large value $\approx 0.065 = \sum_\alpha q_\alpha^2$	

$C_1 \; C_2 \dots C_N$
$\parallel \quad \parallel$
$C_1 \; C_2 \; C_3 \dots C_N$

recover t.

Why? Assume pl.-text is randomly chosen.

i fixed

$$P_r(C_i = C_{i+d}) = P_r(\underbrace{P_u(m_i)}_{\text{"}c_i} = P_v(\underbrace{m_{i+d}}_{c_{i+d}})) =$$

$u \equiv i \bmod t$
$v \equiv i+d \bmod t$

$$= P_r(m_i = \underbrace{P_u^{-1} P_v}_{\text{"}R}(m_{i+d})) =$$

$$= \sum_{\alpha \in A} P_r(m_i = R(m_{i+d}), m_{i+d} = \alpha)$$

complete probability formula

$$= \sum_\alpha P_r(m_i = R(\alpha), m_{i+d} = \alpha) \approx$$

joint even

$$\approx \sum_\alpha P_r(m_i = R(\alpha)) \cdot P_r(m_{i+d} = \alpha)$$

By independence of m_i, m_{i+d} for rel. large d.

$$= \sum_\alpha q_{R(\alpha)} \cdot q_\alpha$$

1) $d \equiv 0 \bmod t \Rightarrow u \equiv v \bmod t \Rightarrow P_u = P_v \Rightarrow R = 1$

↓ for $R = 1$

$$P_r(c_i = c_{i+d}) = \sum_\alpha q_\alpha^2 \approx 0.065.$$

2) $d \not\equiv 0 \bmod t \Rightarrow R$ looks as a random subst.

average of $\sum_\alpha q_{R(\alpha)} \cdot q_\alpha$ over all subs. R.

$$\frac{1}{26!} \sum_R \sum_\alpha q_{R(\alpha)} \cdot q_\alpha =$$

$$= \frac{1}{26!} \sum_\alpha q_\alpha \sum_R q_{R(\alpha)} = \frac{25!}{26!} \sum_\alpha q_\alpha^2 = \frac{1}{26} \approx 0.038$$

$$\sum_R q_{R(\alpha)} = \sum_{\beta \in A} \sum_{R:\, R(\alpha)=\beta} q_\beta =$$

$$= \sum_\beta q_\beta \sum_{R:\, R(\alpha)=\beta} 1 = 25! \sum_\beta q_\beta = 25!$$

$0 \ldots \alpha \ldots 25$ — $25!$

$\ldots \beta \ldots$

We now know t. How to find pe-text? $c_1 c_2 \ldots c_N$

write cipher-text as an array $t \times \approx N/t$

$\approx N/t$ P_i

$C(1)$	$c_1\ c_{1+t}\ c_{1+2t} \ldots$	$\leftarrow$	$m_1\ m_{1+t}\ m_{1+2t} \ldots$
$C(2)$	$c_2\ c_{2+t}\ c_{2+2t} \ldots$	$\leftarrow P_2$	$m_2\ m_{2+t}\ m_{2+2t} \ldots$
$\ldots$			
$C(t)$	$c_t\ c_{2t}\ c_{3t} \ldots$	$\leftarrow P_t$	$m_t\ m_{2t}\ m_{3t} \ldots$

look at $C(i)$

Find frequencies of $cR.$ in $C(i)$

$r_0 r_1 \dots r_{25} = r_a r_b \dots r_z$

$r_{P_i(d)} \approx q_d$ frequency of $d \in A$ in long English texts.

1 Full Vigenère $P_1 \dots P_t$ are random subs. on A.

guess on encryption of E, T, A, ...

do that for all $C(i)$, partly decrypt cipher-text and check the guesses.

But if N is large enough, then it works.

$N \approx 100 \cdot t$

2. simple Vigenère

$$P_i(d) = d + k_i \bmod 26$$

A B C ... Z
0 1 2 ... 25

take row $C(i)$, compute frequencies in $C(i)$

$r_0 r_1 \dots r_{25}$

$$r_{d + k_i \bmod 26} \approx q_d$$

measure for possible shifts s:

$$M_s = \sum_d q_d \cdot r_{d+s \bmod 26}$$

two cases

1) $s = k_i$ correct shift

$$M_s = \sum_d q_d r_{d+k_i} \approx \sum_d q_d \cdot q_d \approx 0.065$$

2) $s \neq k_i$

average of M_s (over all shifts s)

$$\approx \frac{1}{26} = 0.038$$

correct shift is recovered by computing

26 values M_s.

Complexity of the attack.

$$t\left(\frac{N}{t} + 26^2\right)$$

$\frac{N}{t}$ ↑ find frequencies in $C(i)$

26^2 ↑ compute 26 of M_s sum 26 terms

$= N + t \cdot 26^2$ some operations

compared with brute force attack $P_1 P_2 \cdots P_t$ 26^t trials.

conclude efficient cryptanalysis.

Example. Lecture Notes

cipher-text

kpRpw qzxrg ---

of length 332 characters.

frequencies

l	0.081
p	0.078
R	0.069
x	0.066

0.137--

assume a Vigenère cipher

Find block size t

apply Kasiski test

7-ch. segments "aaRjeyx" in positions 309, 209 in the c-text.

distance 309-209 = 100

"

3-ch. "psx" in positions 258, 73

distance $= 258 - 73 = 185$

$\gcd(100, 185) = 5$

there are a lot of 3-ch. identical segments in the cipher-text at distances multiple of 5

$\Rightarrow \quad t = 5.$

array of 5 rows and $\frac{332}{5} \approx 66$

$C(1) = kqzzk\ldots$ of length 67

frequencies in $C(1)$

$z = (\underset{a}{0}, \underset{b}{0}, \underset{c}{0}, \underset{d}{3/67}, 2/67, 3/67, 2/67, 12/67, \ldots)$

compute $(M_0, M_1, \ldots, M_{25}) =$

$(\underset{0}{0.032}, \underset{1}{0.031}, \underset{2}{0.041}, 0.065, 0.043, \ldots)$

$\approx \sum_{\alpha} q_\alpha^2$

$\Rightarrow \quad k_1 = 3$

similar for $C(2), \ldots, C(5)$

recover cipher key $(3, 11, 4, 7, 19)$

decrypt cipher-text to

HE DID NOT KNOW...

Index of Coincidence.

$x = x_1 x_2 \ldots x_N$ N-ch. string over Latin alphabet A.

index $I_c(x)$ measures how close x to a sensible English text.

$I_c(x)$ probability that

definition. $I_c(x)$ / coincidence index

two random x_i, x_j from x $(i \neq j)$ are identical $x_i = x_j$.

introduce indicator

$$\gamma_{ij} = \begin{cases} 1 & x_i = x_j \\ 0 & x_i \neq x_j \end{cases} \qquad 1 \leq i < j \leq N$$

$$I_c(x) = \frac{1}{\binom{N}{2}} \cdot \sum_{1 \leq i < j \leq N} \gamma_{ij}$$

$$\binom{N}{2} = \frac{N(N-1)}{2}$$

find expectation of $I_c(x)$

1) x random string of characters
2) $x = m$ random English text.

then we prove a Baudy formula to compute $I_c(x)$.

1) x random N-character string.

$$E\, I_c(x) = \frac{1}{\binom{N}{2}} \sum_{1 \leq i < j \leq N} E\, \gamma_{ij}$$

↑ expectation is a linear operator

$$E\, \gamma_{ij} = 1 \cdot Pr(x_i = x_j) + 0 \cdot Pr(x_i \neq x_j) =$$
$$= P_2(x_i = x_j)$$

$$\Rightarrow E\, I_c(x) = \frac{1}{\binom{N}{2}} \sum_{1 \leq i < j \leq N} P_2(x_i = x_j) = *$$

$$P_2(x_i = x_j) = \frac{1}{26}$$

$$* = \frac{1}{\binom{N}{2}} \cdot \binom{N}{2} \cdot \frac{1}{26} = \frac{1}{26} \approx \underline{0.038}$$

2) $x = m$ random English text

$$E\, I_c(m) = \frac{1}{\binom{N}{2}} \sum_{1 \le i < j \le N} P_2(m_i = m_j)$$

$$P_2(m_i = m_j) = \sum_{\alpha \in A} P_2(m_i = \alpha, m_j = \alpha) =$$

$$= \sum_{\alpha} P_2(m_i = \alpha, m_j = \alpha) = \quad (j = i+1,\ m_i, m_{i+1}\ \text{IH})$$

$$\approx \sum_{\alpha} P_2(m_i = \alpha) \cdot P_2(m_j = \alpha) = \sum_{\alpha} q_\alpha^2 \approx 0.065$$

($P_2(m_i = \alpha) \sim q_\alpha$, $P_2(m_j = \alpha) \sim q_\alpha$)

$$\Rightarrow E\, I_c(m) \approx \underline{0.065}$$

Why do we need $I_c(x)$?

automate decryption.

formula to compute $I_c(x)$

$x = x_1 \dots x_N$

$f_\alpha = \#\ \alpha$ appears in x.

$$I_c(x) = \frac{1}{\binom{N}{2}} \sum_{\alpha \in A} \binom{f_\alpha}{2} = \frac{\sum_{\alpha} f_\alpha (f_\alpha - 1)}{N(N-1)}$$

$\binom{f_\alpha}{2} = \#\ x_i = x_j,\ i \neq j$

apply to text

x = THE DATA ENCRYPTION STANDARD

α	p_α
A, T	4
D, N	3
E, R	2
C, H, I, O, S, Y	1

$$\Rightarrow I_c(x) = \frac{\overset{A}{4}\cdot\overset{T}{3} + \overset{D}{3}\cdot\overset{N}{2} + \overset{E}{2}\cdot\overset{R}{1} + 1\cdot 0}{25\cdot 24} =$$

$$= \frac{40}{600} \approx 0.066 \simeq \sum_\alpha q_\alpha^2$$

$\Rightarrow$ x is likely to be a sensible English text.

index may be defined for digrams, trigrams, ...