RWTH AACHEN UNIVERSITY

## 3.3. Estimating the key length of Vigenère ciphers

Stochastic model:

$\mathcal{X} = \{0, \dots, m-1\}$ Alphabet

$k$ keyword length, $n$ message length, $k|n$

$$M = (M_1, \dots, M_k, M_{k+1}, \dots, M_{2k}, M_{2k+1}, \dots, M_n)$$

$$+ \quad K = (K_1, \dots, K_k, K_1, \dots, K_k, K_1, \dots, K_k)$$

$$C = (C_1, \dots, C_k, C_{k+1}, \dots, C_{2k}, C_{2k+1}, \dots, C_n)$$

$M_i$ i.i.d., $P(M_i = \ell) = p_\ell$ (known)

$K_i$ i.i.d., $P(K_i = \ell) = \frac{1}{m}$

$I_C$ : index of coincidence,

$$I_C = \frac{1}{\binom{n}{2}} \sum_{i<j} Y_{ij}, \quad Y_{ij} = \begin{cases} 1, & C_i = C_j \\ 0, & \text{otherwise} \end{cases} \quad 1 \le i < j \le n$$

$$\kappa_M = \sum_{\ell=0}^{m-1} p_\ell^2$$

**Lemma 3.5.** $$E(I_C) = \frac{1}{k(n-1)} \left[ (n-k)\kappa_M + n(k-1)\frac{1}{m} \right] \quad (*)$$

Thinking the Future
Zukunft denken

RWTH AACHEN UNIVERSITY

Outline of the proof.

Consider 2 cases:

1.) $i \equiv j \pmod k$

$$E(Y_{ij}) = \sum_{\ell=0}^{m-1} p_\ell^2 = \kappa_M$$

2.) $i \not\equiv j \pmod k$

$$E(Y_{ij}) = \frac{1}{m}$$

Finally: $E(I_c) = \frac{1}{\binom{n}{2}} \sum_{i<j} E(Y_{ij})$

$$= \frac{1}{\binom{n}{2}} \Big[ \sum_{\substack{i<j \\ i\sim j}} E(Y_{ij}) + \sum_{\substack{i<j \\ i\not\sim j}} E(Y_{ij}) \Big]$$

$= (*)$

We are interested in k. Solve $(*)$ for $k$:

$$(n-1)\, E(I_c) = \frac{1}{k}\left( n\left(\kappa_M - \frac{1}{m}\right) - \left(\kappa_M - \frac{n}{m}\right)\right)$$

$$k = \frac{n\left(\kappa_M - \frac{1}{m}\right)}{(n-1)\,E(I_c) + \kappa_M - \frac{n}{m}}$$

Thinking the Future
Zukunft denken

Application: Estimate $E(I_c)$ by $I_c$

$$I_c = \frac{1}{n(n-1)} \sum_{\ell=1}^{n} n_\ell (n_\ell - 1)$$

By Lemma 3.3. ff.: $I_c \longrightarrow E(I_c)$ $(n \to \infty)$ a.e.

In German: $\kappa_M = 0.0762$, $n = 26$

Hence:
$$\hat{k} = \frac{0.0377n}{(n-1)I_c - 0.0385n + 0.0762}$$

If $k$ is known, write $C$ as follows

$$\tilde{C} = \begin{pmatrix} c_1 & \cdots & c_k \\ c_{k+1} & \cdots & c_{2k} \\ \vdots & & \vdots \\ c_{sk+1} & \cdots & c_n \end{pmatrix}$$

The columns are monoalphabetic, apply frequency analysis to the columns.

# 3.4. Vigenère cipher with running key

$$\begin{array}{rcccc} & a_1 & a_2 & \cdots & a_n \\ \oplus & s_1 & s_2 & \cdots & s_n \\ \hline & c_1 & c_2 & \cdots & c_n \end{array}$$ (taken from a book)

Frequency attack is possible, if $(s_1, \dots, s_n)$ is from a nat. language.

Model: $M_i$ r.v.s occurrence of plaintext char
$K_i$ r.v.s " " key char. $\Big\}$ stoch. ind.

Consider most frequ. char.: E, T, A, O, I, N, S (57%)

$P((M_i, K_i) \in \{E, \dots, S\}^2) = 0.57^2 = 0.3249$

About $\frac{1}{3}$ of all ciphert. char. are obtained by "adding" 2 of the most frequent char.

*Most frequent characters:* *(total 57.29%)*
*E - 12.51%, T - 9.25%, A - 8.04%, O - 7.60%, I - 7.26%, N - 7.09%, S - 6.54%*

*Example:*

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

*Investigating the first five ciphertext characters:*

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

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

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

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

There are $3 \cdot 3 \cdot 4 \cdot 2 = 72$ pairs out of the most frequent characters.

Some of them are:

EIAAN $\cdots$ ITAAN $\cdots$ ITEIN $\cdots$ ITINN $\cdots$ ITISN
IEIAS $\cdots$ ETIAS $\cdots$ ETESS $\cdots$ ETANS $\cdots$ ETAIS

Defense against this attack: random key stream

→ one time pad

However, never use the same key twice. Otherwise:

$$(a_1, \dots, a_4) \oplus (k_1, \dots, k_4) = (c_1, \dots, c_4)$$
$$(b_1, \dots, b_4) \oplus (k_1, \dots, k_4) = (d_1, \dots, d_4)$$

Oscar:

$$(c_i - d_i) \bmod 26 = (a_i - b_i) \bmod 26$$

vulnerable to the above attack.

# 4. Entropy and Perfect Secrecy

## 4.1. Entropy

Consider random experiments, e.g.,

$(0.9, 0.05, 0.05)$

$(0.33, 0.33, 0.34)$

We aim at a measure of

$$\overset{!}{=} \begin{cases} \text{uncertainty about the outcome (before)} \\ \text{information gained by the outcome (after)} \end{cases}$$

The right measure was introduced by Shannon (1949).

Formal description

$X$ : discrete r.v. with finite support $\mathcal{X} = \{x_1, \dots, x_m\}$

distribution : $P(X=x_i) = p_i, \; i = 1, \dots, m$

<u>Def. 4.1.</u> Let $c > 1$ constant.

$$H(X) = -\sum_{i=1}^{m} p_i \log_c p_i = -\sum_{i=1}^{m} P(X=x_i) \log_c P(X=x_i)$$

is called <u>entropy</u> of $X$ (or $(p_1, \dots, p_m)$). $\lrcorner$

Convention: $0 \cdot \log 0 = 0$, omit $c$ but fix it.

Thinking the Future
Zukunft denken

Analogously for 2-dim. random variables

$(X,Y)$ with support $\mathcal{X} \times \mathcal{Y} = \{x_1, \dots, x_m\} \times \{y_1, \dots, y_d\}$

distribution $P(X=x_i, Y=y_j) = p_{ij}$

Def 4.2.

a) $$H(X,Y) = -\sum_{i,j} P(X=x_i, Y=y_j) \log P(X=x_i, Y=y_j)$$
$$= -\sum_{i,j} p_{ij} \log p_{ij}$$

is called (joint) entropy of $X, Y$.

b) $$H(X|Y) = -\sum_{j=1}^{d} P(Y=y_j) \sum_{i=1}^{m} P(X=x_i | Y=y_j) \log P(X=x_i | Y=y_j)$$
$$= -\sum_{i,j} P(X=x_i, Y=y_j) \log P(X=x_i | Y=y_j)$$

is called conditional entropy or equivocation. $\perp$

RWTH AACHEN UNIVERSITY

Theorem 4.3.

a) $0 \overset{(i)}{\leq} H(X) \overset{(ii)}{\leq} \log m$

"=" in (i) $\Leftrightarrow$ $\exists x_i: P(X=x_i)=1$

"=" in (ii) $\Leftrightarrow$ $P(X=x_i)=\frac{1}{m} \quad \forall i$

b) $0 \overset{(i)}{\leq} H(X|Y) \overset{(ii)}{\leq} H(X)$

"=" in (i) $\Leftrightarrow$ $P(X=x_i|Y=y_j)=1 \quad \forall i,j$ with $P(X=x_i, Y=y_j)>0$

"=" in (ii) $\Leftrightarrow$ X, Y stoch. indep.

c) $H(X) \overset{(i)}{\leq} H(X,Y) \overset{(ii)}{\leq} H(X)+H(Y)$

"=" in (i) $\Leftrightarrow$ Y is fully dependendent on X

"=" in (ii) $\Leftrightarrow$ X, Y stoch. independent

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

(chain rule)

$\perp$

Proof. any book on information theory

Thinking the Future
Zukunft denken