7 Coincidences of Two Texts

The first six sections of this chapter introduced efficient methods for recognizing plaintext in comparison with noise. These methods break down for encrypted texts because they ignore properties that remain invariant under encryption. One such invariant property—at least for monoalphabetic substitution—is the equality of two letters, no matter what the concrete value of these letters is.

This is the main idea that we work out in the next sections: Look for identical letters in one or more texts, or in other words, for coincidences.

Definition

Let Σ be a finite alphabet. Let $a = (a_0, \ldots, a_{r-1})$ and $b = (b_0, \ldots, b_{r-1}) \in \Sigma^r$ be two texts of the same length $r \geq 1$. Then

$$\kappa(a,b) := \frac{1}{r} \cdot \#\{j \mid a_j = b_j\} = \frac{1}{r} \cdot \sum_{j=0}^{r-1} \delta_{a_j b_j}$$

is called **coincidence index** of a and b (where δ = Kronecker symbol).

For each $r \in \mathbb{N}_1$ this defines a map

$$\kappa\colon \Sigma^r \times \Sigma^r \longrightarrow \mathbb{Q} \subseteq \mathbb{R}.$$

The scaling factor $\frac{1}{r}$ makes results for different lengths comparable.

A Perl program is in the Web: `http://www.staff.uni-mainz.de/pommeren/Cryptology/Classic/Perl/kappa.pl`.

Remarks

1. Always $0 \leq \kappa(a,b) \leq 1$.
2. $\kappa(a,b) = 1 \Longleftrightarrow a = b$.
3. By convention $\kappa(\emptyset,\emptyset) = 1$ (where $\emptyset$ denotes the empty string by abuse of notation).
4. Note that up to scaling the coincidence index is a converse of the Hamming distance that counts non-coincidences.

Example 1: Two English Texts

We compare the first four verses (text 1) of the poem “If ...” by Rudyard Kipling and the next four verses (text 2). (The lengths differ, so we crop the longer one.)

```
IFYOU CANKE EPYOU RHEAD WHENA LLABO UTYOU ARELO OSING THEIR
IFYOU CANMA KEONE HEAPO FALLY OURWI NNING SANDR ISKIT ONONE
||||| |||                                        |
SANDB LAMIN GITON YOUIF YOUCA NTRUS TYOUR SELFW HENAL LMEND
TURNO FPITC HANDT OSSAN DLOOS EANDS TARTA GAINA TYOUR BEGIN
                                  | |
OUBTY OUBUT MAKEA LLOWA NCEFO RTHEI RDOUB TINGT OOIFY OUCAN
NINGS ANDNE VERBR EATHE AWORD ABOUT YOURL OSSIF YOUCA NFORC
                                                 |

WAITA NDNOT BETIR EDBYW AITIN GORBE INGLI EDABO UTDON TDEAL
EYOUR HEART ANDNE RVEAN DSINE WTOSE RVEYO URTUR NLONG AFTER
          |                       |
INLIE SORBE INGHA TEDDO NTGIV EWAYT OHATI NGAND YETDO NTLOO
THEYA REGON EANDS OHOLD ONWHE NTHER EISNO THING INYOU EXCEP
                                             |
KTOOG OODNO RTALK TOOWI SEIFY OUCAN DREAM ANDNO TMAKE DREAM
TTHEW ILLWH ICHSA YSTOT HEMHO LDONI FYOUC ANTAL KWITH CROWD
 |                       |                ||           |
SYOUR MASTE RIFYO UCANT HINKA NDNOT MAKET HOUGH TSYOU RAIMI
SANDK EEPYO URVIR TUEOR WALKW ITHKI NGSNO RLOOS ETHEC OMMON
|                          |
FYOUC ANMEE TWITH TRIUM PHAND DISAS TERAN DTREA TTHOS ETWOI
TOUCH IFNEI THERF OESNO RLOVI NGFRI ENDSC ANHUR TYOUI FALLM
         |  |                                   |
MPOST ORSAS THESA MEIFY OUCAN BEART OHEAR THETR UTHYO UVESP
ENCOU NTWOR THYOU BUTNO NETOO MUCHI FYOUC ANFIL LTHEU NFORG
            ||                                   ||
OKENT WISTE DBYKN AVEST OMAKE ATRAP FORFO OLSOR WATCH THETH
IVING MINUT EWITH SIXTY SECON DSWOR THOFD ISTAN CERUN YOURS
   |   |                               |
INGSY OUGAV EYOUR LIFEF ORBRO KENAN DSTOO PANDB UILDE MUPWI
ISTHE EARTH ANDEV ERYTH INGTH ATSIN ITAND WHICH ISMOR EYOUL
|                                 |
THWOR NOUTT OOLS
LBEAM ANMYS ON
            |
```

In these texts of length 562 we find 35 coincidences, the coincidence index is $\frac{35}{562} = 0.0623$.

Invariance

The coincidence index of two texts is an invariant of polyalphabetic substitution (the keys being equal):

Proposition 1 (Invariance) *Let $f\colon \Sigma^* \longrightarrow \Sigma^*$ be a polyalphabetic encryption function. Then*

$$\kappa(f(a), f(b)) = \kappa(a, b)$$

for all $a, b \in \Sigma^*$ of the same length.

Note that Proposition 1 doesn't need any assumptions on periodicity or on relations between the alphabets used. It only assumes that the encryption function uses the same alphabets at the corresponding positions in the texts.

Mean Values

For a fixed $a \in \Sigma^r$ we determine the mean value of $\kappa(a, b)$ taken over all $b \in \Sigma^r$:

$$\begin{aligned}
\frac{1}{n^r} \cdot \sum_{b \in \Sigma^r} \kappa(a, b) &= \frac{1}{n^r} \cdot \sum_{b \in \Sigma^r} \left[\frac{1}{r} \cdot \sum_{j=0}^{r-1} \delta_{a_j b_j} \right] \\
&= \frac{1}{rn^r} \cdot \sum_{j=0}^{r-1} \underbrace{\left[\sum_{b \in \Sigma^r} \delta_{a_j b_j} \right]}_{n^{r-1}} \\
&= \frac{1}{rn^r} \cdot r \cdot n^{r-1} = \frac{1}{n},
\end{aligned}$$

because, if $b_j = a_j$ is fixed, there remain n^{r-1} possible values for b.

In an analogous way we determine the mean value of $\kappa(a, f_\sigma(b)$ for fixed $a, b \in \Sigma^r$ over all permutations $\sigma \in \mathcal{S}(\Sigma)$:

$$\begin{aligned}
\frac{1}{n!} \cdot \sum_{\sigma \in \mathcal{S}(\Sigma)} \kappa(a, f_\sigma(b)) &= \frac{1}{n!} \cdot \frac{1}{r} \sum_{\sigma \in \mathcal{S}(\Sigma)} \#\{j \mid \sigma b_j = a_j\} \\
&= \frac{1}{rn!} \cdot \#\{(j, \sigma) \mid \sigma b_j = a_j\} \\
&= \frac{1}{rn!} \cdot \sum_{j=0}^{r-1} \#\{\sigma \mid \sigma b_j = a_j\} \\
&= \frac{1}{rn!} \cdot r \cdot (n-1)! = \frac{1}{n},
\end{aligned}$$

because exactly $(n-1)!$ permutations map a_j to b_j.

Note that this conclusion also works for $a = b$.

This derivation shows:

Proposition 2 (i) *The mean value of $\kappa(a,b)$ over all texts $b \in \Sigma^*$ of equal length is $\frac{1}{n}$ for all $a \in \Sigma^*$.*

(ii) *The mean value of $\kappa(a,b)$ over all $a, b \in \Sigma^r$ is $\frac{1}{n}$ for all $r \in \mathbb{N}_1$.*

(iii) *The mean value of $\kappa(a, f_\sigma(b))$ over all monoalphabetic substitutions with $\sigma \in \mathcal{S}(\Sigma)$ is $\frac{1}{n}$ for each pair $a, b \in \Sigma^*$ of texts of equal length.*

(iv) *The mean value of $\kappa(f_\sigma(a), f_\tau(b))$ over all pairs of monoalphabetic substitutions, with $\sigma, \tau \in \mathcal{S}(\Sigma)$, is $\frac{1}{n}$ for each pair $a, b \in \Sigma^*$ of texts of equal length.*

Interpretation

- For a given text a and a "random" text b of the same length $\kappa(a,b) \approx \frac{1}{n}$.
- For "random" texts a and b of the same length $\kappa(a,b) \approx \frac{1}{n}$.
- For given texts a and b of the same length and a "random" monoalphabetic substitution f_σ we have $\kappa(a, f_\sigma(b)) \approx \frac{1}{n}$. This remark justifies treating a nontrivially monoalphabetically encrypted text as random with respect to κ and plaintexts.
- For given texts a and b of the same length and two "random" monoalphabetic substitutions f_σ, f_τ we have $\kappa(f_\sigma(a), f_\tau(b)) \approx \frac{1}{n}$.
- The same holds for "random" polyalphabetic substitutions because counting the coincidences is additive with respect to arbitrary decompositions of texts.

Values that significantly differ from these mean values are suspicious for the cryptanalyst, they could have a *non-random cause*. For more precise statements we should assess the variances (or standard deviations) or, more generally, the distribution of κ-values in certain "populations" of texts.

Variance

First fix $a \in \Sigma^r$ and vary b over all of Σ^r. Using the mean value $\frac{1}{n}$ we calculate the variance:

$$\begin{aligned} V_{\Sigma^r}(\kappa, a) &= \frac{1}{n^r} \cdot \sum_{b \in \Sigma^r} \kappa(a,b)^2 - \frac{1}{n^2} \\ &= \frac{1}{n^r} \cdot \sum_{b \in \Sigma^r} \left[\frac{1}{r} \cdot \sum_{j=0}^{r-1} \delta_{a_j b_j} \right]^2 - \frac{1}{n^2} \end{aligned}$$

Evaluating the square of the sum in brackets we get the quadratic terms

$$\sum_{j=0}^{r-1} \delta_{a_j b_j}^2 = \sum_{j=0}^{r-1} \delta_{a_j b_j} = r \cdot \kappa(a,b) \quad \text{because} \quad \delta_{a_j b_j} = 0 \text{ or } 1$$

$$\sum_{b\in\Sigma^r}\sum_{j=0}^{r-1}\delta^2_{a_jb_j} = r\cdot\sum_{b\in\Sigma^r}\kappa(a,b) = r\cdot n^r\cdot\frac{1}{n} = r\cdot n^{r-1}$$

and the mixed terms

$$2\cdot\sum_{j=0}^{r-1}\sum_{k=j+1}^{r-1}\delta_{a_jb_j}\delta_{a_kb_k} \quad\text{where}\quad \delta_{a_jb_j}\delta_{a_kb_k} = \begin{cases}1 & \text{if } a_j=b_j \text{ and } a_k=b_k\\ 0 & \text{else}\end{cases}$$

If we fix two letters b_j and b_k, we are left with n^{r-2} different b's that give the value 1. The total sum over the mixed terms evaluates as

$$\sum_{b\in\Sigma^r}\left(2\cdot\sum_{j=0}^{r-1}\sum_{k=j+1}^{r-1}\delta_{a_jb_j}\delta_{a_kb_k}\right) = 2\cdot\sum_{j=0}^{r-1}\sum_{k=j+1}^{r-1}\underbrace{\sum_{b\in\Sigma^r}\delta_{a_jb_j}\delta_{a_kb_k}}_{n^{r-2}}$$

Substituting our intermediary results we get

$$\begin{aligned} V_{\Sigma^r}(\kappa,a) &= \frac{1}{n^rr^2}\left(r\cdot n^{r-1}+r\cdot(r-1)\cdot n^{r-2}\right)-\frac{1}{n^2}\\ &= \frac{1}{rn}+\frac{r-1}{rn^2}-\frac{1}{n^2} = \frac{1}{rn}-\frac{1}{rn^2} = \frac{1}{r}\left(\frac{1}{n}-\frac{1}{n^2}\right)\end{aligned}$$

Next we let a and b vary and calculate the variance of κ:

$$\begin{aligned} V_{\Sigma^r}(\kappa) &= \frac{1}{n^{2r}}\sum_{a,b\in\Sigma^r}\kappa(a,b)^2-\frac{1}{n^2}\\ &= \frac{1}{n^r}\sum_{a\in\Sigma^r}\underbrace{\left(\frac{1}{n^r}\sum_{b\in\Sigma^r}\kappa(a,b)^2\right)}_{\frac{1}{r}\left(\frac{1}{n}-\frac{1}{n^2}\right)+\frac{1}{n^2}}-\frac{1}{n^2}\\ &= \frac{1}{r}\left(\frac{1}{n}-\frac{1}{n^2}\right)+\frac{1}{n^2}-\frac{1}{n^2} = \frac{1}{r}\left(\frac{1}{n}-\frac{1}{n^2}\right)\end{aligned}$$

We have shown:

Proposition 3 (i) *The mean value of $\kappa(a,b)$ over all texts b of equal length $r\in\mathbb{N}_1$ is $\frac{1}{n}$ with variance $\frac{1}{r}\left(\frac{1}{n}-\frac{1}{n^2}\right)$ for all $a\in\Sigma^r$.*

(ii) *The mean value of $\kappa(a,b)$ over all $a,b\in\Sigma^r$ is $\frac{1}{n}$ with variance $\frac{1}{r}\left(\frac{1}{n}-\frac{1}{n^2}\right)$ for all $r\in\mathbb{N}_1$.*

For the 26 letter alphabet A...Z we have the mean value $\frac{1}{26}\approx 0.0385$, independently from the text length r. The variance is $\approx\frac{0.03370}{r}$, the standard deviation $\approx\frac{0.19231}{\sqrt{r}}$. From this we get the second row of Table 20.

For statistical tests (one-sided in this case) we would like to know the 95% quantiles. If we take the values for a normal distribution as approximations,

Table 20: *Standard deviations and 95% quantiles of κ for random text pairs of length r*

r	10	40	100	400	1000	10000
Std dev	0.0608	0.0304	0.0192	0.0096	0.0061	0.0019
95% quantile	0.1385	0.0885	0.0700	0.0543	0.0485	0.0416

that is "mean value + 1.645 times standard deviation", we get the values in the third row of Table 20. These raw estimates show that the κ-statistic in this form is weak in distinguishing "meaningful" texts from random texts, even for text lengths of 100 letters, and strong only for texts of several thousand letters.

Distinguishing meaningful plaintext from random noise is evidently not the main application of the κ-statistic. The next section will show the true relevancy of the coincidence index.