9 Autoincidence of a Text

Introduction

For the cryptanalysis of periodic polyalphabetic ciphers the following construction is of special importance: Let $a \in \Sigma^*$, and let $a_{(q)}$ and $a_{(-q)}$ be the cyclic shifts of a by q positions to the right resp. to the left. That is

$$\begin{array}{lcllllllllll} a & = & a_0 & a_1 & a_2 & \ldots & a_{q-1} & a_q & a_{q+1} & \ldots & a_{r-1} \\ a_{(q)} & = & a_{r-q} & a_{r-q+1} & a_{r-q+2} & \ldots & a_{r-1} & a_0 & a_1 & \ldots & a_{r-q-1} \\ a_{(-q)} & = & a_q & a_{q+1} & a_{q+2} & \ldots & a_{2q-1} & a_{2q} & a_{2q+1} & \ldots & a_{q-1} \end{array}$$

Clearly $\kappa(a, a_{(q)}) = \kappa(a, a_{(-q)})$.

Definition. For a text $a \in \Sigma^*$ and a natural number $q \in \mathbb{N}$ the number $\kappa_q(a) := \kappa(a, a_{(q)})$ is called the q-th **autocoincidence index** of a.

Note. This is not a common notation. Usually this concept is not given an explicit name.

Example. We shift a text by 6 positions to the right:

```
COINCIDENCESBETWEENTHETEXTANDTHESHIFTEDTEXT <-- original text
EDTEXTCOINCIDENCESBETWEENTHETEXTANDTHESHIFT <-- shifted by 6
             |  |      | |           |    | <-- 6 coincidences
```

Properties

The q-th autocoincidence index κ_q defines a map

$$\kappa_q \colon \Sigma^* \longrightarrow \mathbb{Q}.$$

Clearly $\kappa_q(a) = \kappa_{r-q}(a)$ for $a \in \Sigma^r$ and $0 < q < r$, and κ_0 is a constant map.

Application

Take a ciphertext c that is generated by a periodic polyalphabetic substitution. If we determine $\kappa_q(c)$, we encounter two different situations: In the general case q is not a multiple of the period l. Counting the coincidences we encounter letter pairs that come from independent monoalphabetic substitutions. By the results of Section 7 we expect an index $\kappa_q(c) \approx \frac{1}{n}$.

In the special case where $l|q$ however we encounter the situation

$$\begin{array}{lllllll} \sigma_0 a_0 & \sigma_1 a_1 & \ldots & \sigma_0 a_q & \sigma_1 a_{q+1} & \ldots \\ & & & \sigma_0 a_0 & \sigma_1 a_1 & \ldots \end{array}$$

where the letters below each other come from the same monoalphabetic substitution. Therefore they coincide if and only if the corresponding plaintext letters coincide. Therefore we expect an index $\kappa_q(c)$ near the coincidence index κ_M that is typical for the plaintext language M.

More precisely for a polyalphabetic substitution f of period l, plaintext a, and ciphertext $c = f(a)$:

1. For l not a divisor of q or $r - q$ we expect $\kappa_q(c) \approx \frac{1}{n}$.
2. For $l|q$ and q small compared with r we expect $\kappa_q(c) \approx \kappa_q(a)$, and this value should be near the typical coincidence index κ_M.

This is the second application of coincidence counts, detecting the period of a polyalphabetic substitution by looking at the autocoincidence indices of the ciphertext. Compared with the search for repetitions after Kasiski this method also takes account of repetitions of length 1 or 2. In this way we make much more economical use of the traces that the period leaves in the ciphertext.

Example

We want to apply these considerations to the autocoincidence analysis of a polyalphabetic ciphertext using the Perl program `coinc.pl` from `http://www.staff.uni-mainz.de/pommeren/Cryptology/Classic/Perl/`. We start with the cryptogram that we already have solved in Chapter 2 by repetition analysis:

```
        00    05    10    15    20    25    30    35    40    45

0000  AOWBK NLRMG EAMYC ZSFJO IYYVS HYQPY KSONE MDUKE MVEMP JBBOA
0050  YUHCB HZPYW MOOKQ VZEAH RMVVP JOWHR JRMWK MHCMM OHFSE GOWZK
0100  IKCRV LAQDX MWRMH XGTHX MXNBY RTAHJ UALRA PCOBJ TCYJA BBMDU
0150  HCQNY NGKLA WYNRJ BRVRZ IDXTV LPUEL AIMIK MKAQT MVBCB WVYUX
0200  KQXYZ NFPGL CHOSO NTMCM JPMLR JIKPO RBSIA OZZZC YPOBJ ZNNJP
0250  UBKCO WAHOO JUWOB CLQAW CYTKM HFPGL KMGKH AHTYG VKBSK LRVOQ
0300  VOEQW EALTM HKOBN CMVKO BJUPA XFAVK NKJAB VKNXX IJVOP YWMWQ
0350  MZRFB UEVYU ZOORB SIAOV VLNUK EMVYY VMSNT UHIWZ WSYPG KAAIY
0400  NQKLZ ZZMGK OYXAO KJBZV LAQZQ AIRMV UKVJO CUKCW YEALJ ZCVKJ
0450  GJOVV WMVCO ZZZPY WMWQM ZUKRE IWIPX BAHZV NHJSJ ZNSXP YHRMG
0500  KUOMY PUELA IZAMC AEWOD QCHEW OAQZQ OETHG ZHAWU NRIAA QYKWX
0550  EJVUF UZSBL RNYDX QZMNY AONYT AUDXA WYHUH OBOYN QJFVH SVGZH
0600  RVOFQ JISVZ JGJME VEHGD XSVKF UKXMV LXQEO NWYNK VOMWV YUZON
0650  JUPAX FANYN VJPOR BSIAO XIYYA JETJT FQKUZ ZZMGK UOMYK IZGAW
0700  KNRJP AIOFU KFAHV MVXKD BMDUK XOMYN KVOXH YPYWM WQMZU EOYVZ
0750  FUJAB YMGDV BGVZJ WNCWY VMHZO MOYVU WKYLR MDJPV JOCUK QELKM
```

```
0800  AJBOS YXQMC AQTYA SABBY ZICOB XMZUK POOUM HEAUE WQUDX TVZCG
0850  JJMVP MHJAB VZSUM CAQTY AJPRV ZINUO NYLMQ KLVHS VUKCW YPAQJ
0900  ABVLM GKUOM YKIZG AVLZU VIJVZ OGJMO WVAKH CUEYN MXPBQ YZVJP
0950  QHYVG JBORB SIAOZ HYZUV PASMF UKFOW QKIZG ASMMK ZAUEW YNJAB
1000  VWEYK GNVRM VUAAQ XQHXK GVZHU VIJOY ZPJBB OOQPE OBLKM DVONV
1050  KNUJA BBMDU HCQNY PQJBA HZMIB HWVTH UGCTV ZDIKG OWAMV GKBBK
1100  KMEAB HQISG ODHZY UWOBR ZJAJE TJTFU K
```

The Autocoincidence Indices

This is the sequence of autocoincidence indices of our cryptogram

κ_1	κ_2	κ_3	κ_4	κ_5	κ_6	κ_7	κ_8
0.0301	0.0345	0.0469	0.0354	0.0371	0.0354	**0.0822**	0.0416
κ_9	κ_{10}	κ_{11}	κ_{12}	κ_{13}	κ_{14}	κ_{15}	κ_{16}
0.0265	0.0309	0.0416	0.0389	0.0327	**0.0787**	0.0460	0.0345
κ_{17}	κ_{18}	κ_{19}	κ_{20}	κ_{21}	κ_{22}	κ_{23}	κ_{24}
0.0460	0.0309	0.0327	0.0309	**0.0769**	0.0318	0.0309	0.0327
κ_{25}	κ_{26}	κ_{27}	κ_{28}	κ_{29}	κ_{30}	κ_{31}	κ_{32}
0.0318	0.0309	0.0416	**0.0875**	0.0477	0.0416	0.0442	0.0354
κ_{33}	κ_{34}	κ_{35}	κ_{36}				
0.0318	0.0389	**0.0610**	0.0371				

The period 7 stands out, as it did with the period analysis after KASISKI in the last chapter. This is also clearly seen in the graphical representation, see Figure 10.

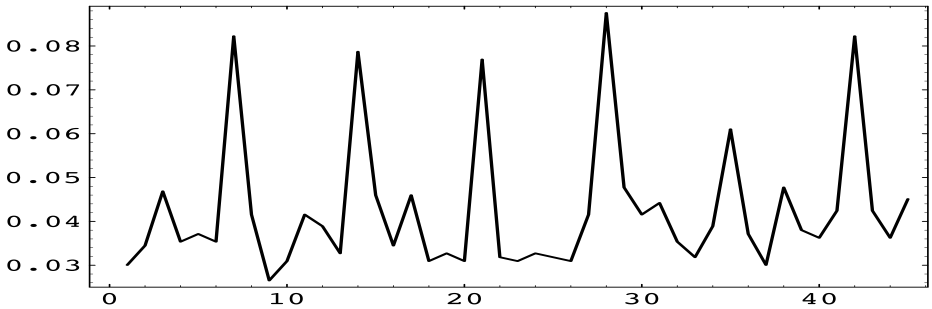

Figure 10: *Autocoincidence spectrum of a sample ciphertext*

The values other than at multiples of 7 fluctuate around the “random” value $\frac{1}{26} \approx 0.0385$ as expected. The values in the peaks fluctuate around the typical coincidence index near 0.08 of the plaintext language German, for which we gave empirical evidence in the last section. This effect has an easy explanation.

The Autocoincidence Spectrum

To analyze the effect seen in Figure 10, let c be the ciphertext from a polyalphabetic encryption of a text $a \in M$ with period l. What values can we expect for the $\kappa_q(c)$?

$c =$	c_0	$\dots$	c_{q-1}	c_q	$\dots$		c_{r-1}
$c_{(q)} =$	c_{r-q}	$\dots$	c_{r-1}	c_0	$\dots$		c_{r-q-1}
expected coinc.:	$q \cdot \kappa_M$	if	$l \mid r-q$,	$(r-q) \cdot \kappa_M$	if		$l \mid q$,
	$q \cdot \kappa_{\Sigma^*}$	else		$(r-q) \cdot \kappa_{\Sigma^*}$	else		

Adding these up we get the following expected values for the autocoincidence spectrum:

1. case, $l \mid r$

$$\kappa_q(c) \approx \begin{cases} \frac{q \cdot \kappa_M + (r-q) \cdot \kappa_M}{r} = \kappa_M & \text{if } l \mid q, \\ \frac{q \cdot \kappa_{\Sigma^*} + (r-q) \cdot \kappa_{\Sigma^*}}{r} = \kappa_{\Sigma^*} & \text{else.} \end{cases}$$

2. case, $l \nmid r$

$$\kappa_q(c) \approx \begin{cases} \frac{q \cdot \kappa_{\Sigma^*} + (r-q) \cdot \kappa_M}{r} & \text{if } l \mid q, \\ \frac{q \cdot \kappa_M + (r-q) \cdot \kappa_{\Sigma^*}}{r} & \text{if } l \mid r-q, \\ \kappa_{\Sigma^*} & \text{else.} \end{cases}$$

In particular for $q << r$

$$\kappa_q(c) \approx \begin{cases} \kappa_M & \text{if } l \mid q, \\ \kappa_{\Sigma^*} & \text{else.} \end{cases}$$

This explains the autocoincidence spectrum that we observed in the example. Typical autocoincidence spectra are shown in Figures 11 and 12.

Since in the second case the resulting image may be somewhat blurred, one could try to calculate autocoincidence indices not by shifting the text cyclically around but by simply cutting off the ends.

Definition. The sequence $(\kappa_1(a), \dots, \kappa_{r-1}(a))$ of autocoincidence indices of a text $a \in \Sigma^r$ of length r is called the **autocoincidence spectrum** of a.

Note. that this notation too is not common in the literature, but seems adequate for its evident cryptanalytical importance.

Exercise 1. Determine the autocoincidence spectrum of the ciphertext that you already broke by a KASISKI analysis. Create a graphical representation of it using graphic software of your choice.

Exercise 2. Cryptanalyze the ciphertext

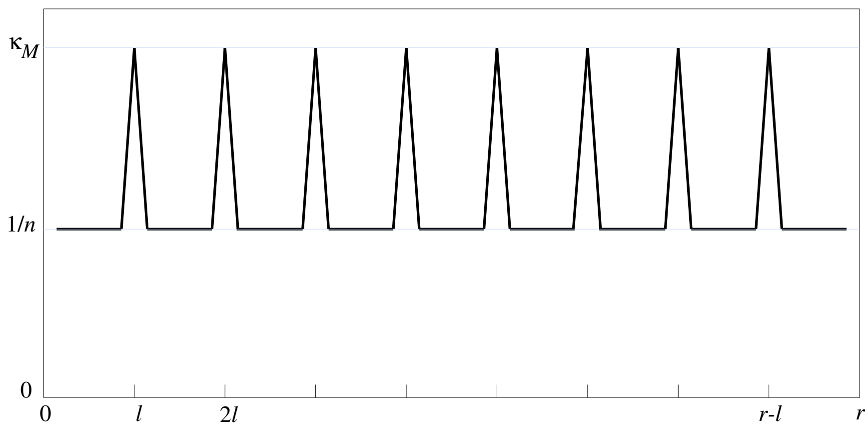

Figure 11: *Text length is multiple of period*

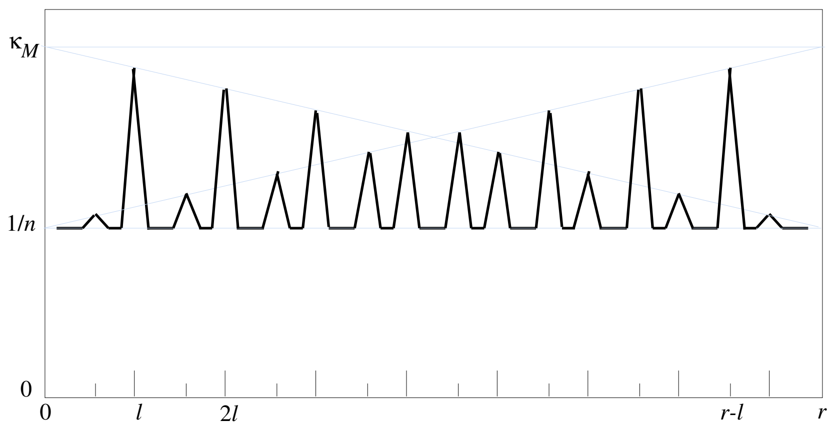

Figure 12: *Text length not multiple of period*

```
ECWUL MVKVR SCLKR IULXP FFXWL SMAEO HYKGA ANVGU GUDNP DBLCK
MYEKJ IMGJH CCUJL SMLGU TXWPN FQAPU EUKUP DBKQO VYTUJ IVWUJ
IYAFL OVAPG VGRYL JNWPK FHCGU TCUJK JYDGB UXWTT BHFKZ UFSWA
FLJGK MCUJR FCLCB DBKEO OUHRP DBVTP UNWPZ ECWUL OVAUZ FHNQY
XYYFL OUFFL SHCTP UCCWL TMWPB OXNKL SNWPZ IIXHP DBSWZ TYJFL
NUMHD JXWTZ QLMEO EYJOP SAWPL IGKQR PGEVL TXWPU AODGA ANZGY
BOKFH TMAEO FCFIH OTXCT PMWUO BOK
```