Tatra Mt. Math. Publ. **45** (2010), 75–91
DOI: 10.2478/v10127-010-0007-8

TATRA MOUNTAINS
Mathematical Publications

THE ZODIAC KILLER CIPHERS

Håvard Raddum — Marek Sýs

ABSTRACT. We describe the background of the Zodiac killer's cipher, and present a strategy for how to attack the unsolved Z340 cipher. We present evidence that there is a high degree of non-randomness in the sequence of ciphertext symbols in this cipher, suggesting it has been constructed in a systematic way. Next, we use this information to design a tool for solving the Zodiac ciphers. Using this tool we are able to re-solve the known Z408 cipher.

1. Background

During the late 1960's and early 1970's there was a serial killer that was active in the San Francisco area in the USA. This serial killer got the name the *Zodiac killer*, and was never caught. There are many theories around the case, and there have been and still are a number of suspects of who the Zodiac killer really was. The last murder that was tied to the Zodiac killer took place in 1974, and after that the series of killings seemed to stop. The official police investigation was closed in 2004.

Most serial killers have some sort of 'signature' on their murders, something that ties the different victims to the same murderer. The signature of the Zodiac killer was that he sent letters to the police and newspapers in San Francisco. Many of these letters contained information and evidence that proved the sender was the killer who had committed the murders.

Some of the letters the Zodiac killer sent contained ciphers. These ciphers look like neatly written rows of symbols. Two of the ciphers contain too little ciphertext to be analyzed, so these have received little attention. These two ciphers were the last ciphers and both of them were sent to the newspaper *San Fransisco Chronicle*, in April and June 1970.

The first ciphertext the Zodiac killer made was divided into three parts, where each piece was sent to the newspapers Vallejo Times-Herald, the San Francisco

2010 Mathematics Subject Classification: Primary: 01A60; Secondary: 6204.
Keywords: classical ciphers.
Supported by the grant NIL-I-004 from Iceland, Liechtenstein and Norway through the EEA Financial Mechanism and the Norwegian Financial Mechanism.

Chronicle and the San Francisco Examiner on July 31st, 1969. He demanded that the newspapers should publish the different parts on their front pages, and all of them did. This cipher was solved in a week, but the plaintext did not reveal any real information. The cipher has been named the *408 cipher*.

The same year, on November 8th, the Zodiac killer sent another cipher to the San Francisco Chronicle. This cipher was also published, but still remains unsolved and has been named the *340 cipher*. In this paper we will examine the 408 and 340 ciphers. We will explain a possible strategy for solving the 340 cipher, and report on the software that was developed for testing this strategy.

2. The ciphers

The Zodiac killer made in total four different ciphers. All of these ciphers are in the form of lines of symbols. At least for the one that was solved, each symbol represents one letter to form a readable English text. We will have to assume, and it seems likely, that the other ciphers also follow the principle of each symbol corresponding to one plaintext letter.

As mentioned above, two of them are too short to be solvable. One of these is included in a letter and should supposedly contain the killer's name. It is a sequence of 13 symbols, of which 8 are different. There are probably several different plausible plaintexts that fit the ciphertext. The other short cipher was written at the bottom of another letter and consists of 32 symbols, of which as many as 29 are different. With this little restriction it is obvious that there are a large number of plaintexts that will match this ciphertext.

2.1. The 408 cipher

This cipher got its name because it contains 408 symbols, of which 54 are different. In all likelihood there can only be one coherent English text that matches the ciphertext. After the newspapers published the parts of the cipher it is probable that many people tried to solve it, and after approximately one week a high school teacher and his wife managed to come up with a solution. This solution must be the correct one, even though the last line does not decipher into meaningful text.

The couple that solved the 408 cipher said they made two assumptions (that afterwards proved to be correct) when they tried to crack it. Several of the letters the Zodiac killer sent were known, and from their content it is easy to infer that the killer had a big ego. They therefore assumed he would start with the word 'I'. The other assumption was that the phrase 'kill' would occur several times in the plaintext. Using this they were able to find the solution by trial and error.

Knowing the solution for the 408 cipher we can look for other clues as for how it has been constructed. Since there are more ciphertext symbols than letters in the English alphabet, some of the ciphertext symbols must represent the same plaintext letter. Which letters that got represented by several symbols, and how many different symbols to use for one letter, is not random. For instance, 'e' being the clearly most common letter in English text is represented by 7 different ciphertext symbols in the 408 cipher. The other most common letters are represented by 3 and 4 symbols each, while the least common letters have a 1–1 correspondence with the symbols encoding them. This indicates that the Zodiac killer was aware of frequency analysis as the tool for solving substitution ciphers, and wanted to actively defend against attacks by frequency analysis.

However, we have observed another source of non-randomness in the 408 cipher. The solution tells us which symbols that map to each plaintext letter. For each letter we can then look at the order of appearance of the symbols representing this letter and write it down. This has been done in Figure 1.

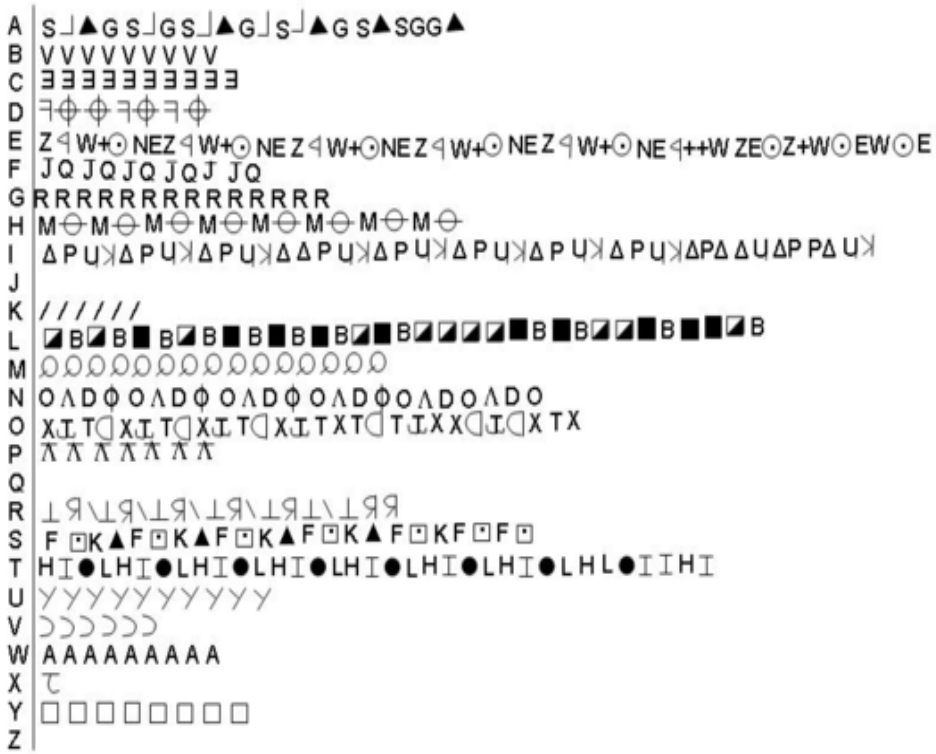

FIGURE 1. For each letter, the symbols representing this letter are written in order of appearance.

We see that the symbols representing the same letter do not appear in random order, but rather in a cyclic order. That is, each symbol is used once before starting over again, reusing the first symbol. We also see that this cycle system has not been followed perfectly, but the tendency towards using the symbols for one letter in order is very clear.

One reason that this strong bias towards a cyclic pattern emerges may be that it is easy to make sure symbols representing the same plaintext letter are used approximately the same number of times. It is important to use these symbols equally often if you want the cipher to be immune to frequency analysis.

2.2. The 340 cipher

Like the 408 cipher, the 340 cipher got its name because it contains 340 symbols. The number of different symbols used in the cipher is 63, so there are more different symbols, and less ciphertext in the 340 cipher compared to the 408. This makes the 340 cipher significantly more difficult to solve, and nobody has been able to come up with a solution, even after 40 years and with computing power that was unimaginable in 1969.

The 340 cipher has received a lot of attention, and there are websites [1], [2] where one can fill in letters for the symbols and try to solve it. The cipher itself, image taken from [3], is shown in Figure 2.

Some people believe that the cipher has never been solved because there is no real plaintext underneath, that the 340 cipher is just a random collection of symbols written to look like a cipher. In the next section we provide evidence against this hypothesis. We believe that the cipher has been constructed in a similar way as the 408 cipher, and that there really is a meaningful plaintext to be found.

3. Looking for cycles in the 340 cipher

The solution for the 408 cipher gives us a hint for how to attack the 340 cipher. Recall that the different symbols representing one plaintext letter in the 408 cipher to a large extent appears in a cyclic order. If the 340 cipher is constructed the same way, the cycle property may give us a way to identify which symbols that represent the same letter (without knowing what the letter is). If we are able to find a small set of ciphertext symbols that appear in a cyclic pattern throughout the cipher, we can guess that these symbols really represent the same plaintext letter. The strategy for attack can then consist of two steps:

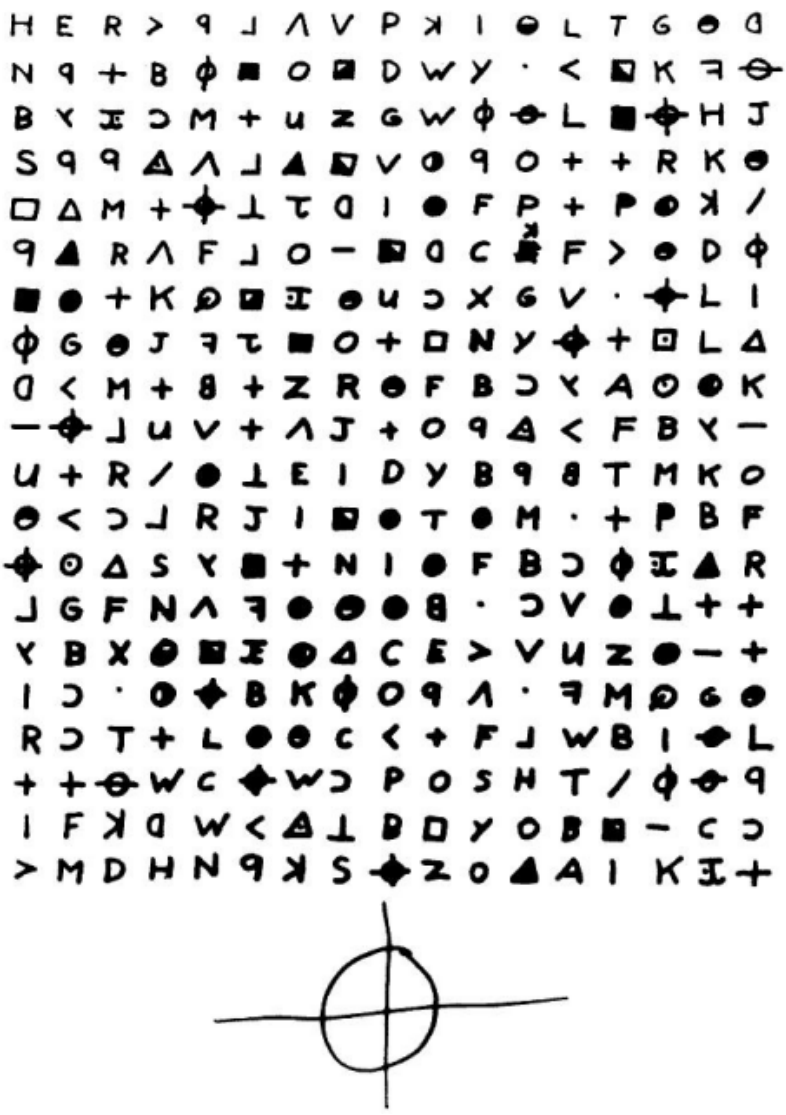

FIGURE 2. The 340 cipher.

- Identify sets of symbols such that the symbols in each set are likely to represent the same letter. Find these sets based on the cycle property.
- Replace the symbols in each set by a single symbol, and use ordinary frequency analysis to find the plaintext.

3.1. Finding 2-cycles

With a computer it is easy to find all sets of two symbols that appear in an alternating manner throughout the whole cipher. This is done by simply trying this property for all $\binom{63}{2}$ pairs of symbols, and recording those pairs that satisfy the cycle property. The number of 2-cycles in the 340 cipher is 90.

3.2. Finding n-cycles

It is fully possible to find the n-cycles by the straight forward exhaustive search approach. After all, $\binom{63}{5} \approx 2^{22.75}$ is not too much for the computer. However, as will be explained later we want to do the cycle count for many different ciphers, so we should look for a faster way. It was not too hard to find a much better algorithm for finding all n-cycles for $n > 2$.

Theorem 1. *Let $s_1, \ldots, s_n$ be ciphertext symbols and $n \geq 3$. Then $s_1, \ldots, s_n$ forms an n-cycle if and only if each of the n sequences $s_2, \ldots, s_n$; $s_1, s_3, \ldots, s_n$; $\ldots$; $s_1, \ldots, s_{n-1}$ are $(n-1)$-cycles.*

Proof.

$\Rightarrow$ Assume the sequence $s_1, \ldots, s_n$ forms an n-cycle in the cipher. Then it is clear that removing one of the symbols s_i will not destroy the cycle property of the others, so each of the n sequences $s_2, \ldots, s_n; \ldots; s_1, \ldots, s_{n-1}$ are $(n-1)$-cycles.

$\Leftarrow$ Assume now that each of the sequences $s_2, \ldots, s_n$; $s_1, s_3, \ldots, s_n$; $\ldots$; $s_1, \ldots, s_{n-1}$ are $(n-1)$-cycles. Assume also that the sequence $s_1, \ldots, s_n$ is *not* an n-cycle. We will prove this yields a contradiction to all the $(n-1)$--subsequences being $(n-1)$-cycles. If $s_1, \ldots, s_n$ is not an n-cycle there must be some symbol s_i that at some point in the ciphertext repeats too quickly, say before s_j has been found. The order of the symbols then looks like $s_i, \ldots, s_i$ at some point in the ciphertext, where s_j does not appear between the s_i. Since $n \geq 3$, there is a symbol s_k that is different to both s_i and s_j. If we remove this s_k, we see that s_i will still repeat too quickly when considering the $(n-1)$--sequence $s_1, \ldots, s_{k-1}, s_{k+1}, \ldots, s_n$, contradicting the base assumption that all subsequences of length $(n-1)$ are $(n-1)$-cycles. Hence $s_1, \ldots, s_n$ must be an n-cycle. □

With Theorem 1 and knowing all the 2-cycles it is easy to recursively find all of the longer cycles that exist in the 340 cipher. For each pair of 2-cycles we will see if there is a common symbol that appears in both of them, so the 2-cycles look like (s_1, s_2) and (s_1, s_3). If s_1 is a common symbol, we will see if (s_2, s_3) also is a 2-cycle. In this case we know that (s_1, s_2, s_3) is a 3-cycle.

In general, after we have found all $(n-1)$-cycles for some n we will look for pairs of $(n-1)$-cycles that overlap in $n-2$ symbols, like $(s_1, \ldots, s_{n-2}, s_{n-1})$ and $(s_1, \ldots, s_{n-2}, s_n)$. When such a pair is found, we will check if each of the

other $n-2$ sequences of length $n-1$ made of symbols from $\{s_1, \ldots, s_n\}$ appear as $(n-1)$-cycles. If they do, we have found an n-cycle. Because Theorem 1 is *if and only if* we know that we will not miss any potential cycles when following this algorithm, and we can thus easily find all cycles that exist in the 340 cipher. The number of cycles for each n is given in Table 1.

TABLE 1. Number of cycles in the 340 cipher.

n	# of cycles
2	90
3	62
4	14
5	2

The number of cycles of various lengths reported above does not really tell us that much by itself. We can guess that the two cycles of length 5 represent one letter each. Then we discard all cycles of shorter length that have some overlap with the 5-cycles. Then we guess that the remaining 4-cycles represent one letter each, and discard all cycles that have some overlap with these, etc. Following this procedure we end up with about 45 sets of symbols, too many for assigning a letter to each.

Moreover, we should not expect this strategy to work since there are many errors in the cycle system in the 408 cipher. Also, many of the cycles we find in the 340 cipher contain symbols that only appear a few times in the cipher, like 2 or 3 times. We should therefore expect that many of the cycles among the symbols with low frequency occur by chance, and do not mean anything.

There are two exceptions to the low-frequency cycles that stand out. The symbol that looks like a square with bottom left half filled with black and the symbol looking like Λ appear 6 times each and forms a 2-cycle. The symbols that look like **M** and **J** appear 7 times each and also form a perfect 2-cycle in the 340 cipher. There are 9 different symbols that appear six times each, and the probability that there is a pair among them that forms a 2-cycle is 43 %. There are only 4 symbols that appear seven times, and the probability that two of them should form a 2-cycle is only 0.5 %. We can not draw any conclusions for the symbols appearing six times, but we find it is quite likely that the symbols **M** and **J** represent the same plaintext letter.

3.3. Cycles in randomized ciphers

We believe that if the cycle system has indeed been used (but with errors), the 340 cipher should contain more cycles than if the cycle system has not been used. We can test this, by writing down the symbols that make up the 340 cipher

in random order, and count the number of cycles in the randomized ciphertext. In the randomized ciphertext there is no cycle system used. We made 10.000 random ciphertexts and counted the number of cycles in each of them. The average number of cycles we find when the cycle system has *not* been used is shown in Table 2.

Table 2. Expected number of cycles in randomized cipher.

n	# of cycles (avg)
2	34.9
3	16.0
4	4.2
5	0.6

We see that the numbers in Tables 1 and 2 differ significantly and that there are many more cycles in the actual 340 cipher than what we should expect if the cycle system has not been used at all when constructing the cipher. We then asked what the maximum number of cycles observed in the 10.000 trials are, and how many instances of the 10.000 that gave numbers higher than those observed in the actual 340 cipher. These numbers are given in Table 3.

Table 3. Most extreme values in 10.000 trials.

n	max. obs.	# times above orig. cipher
2	79	0
3	102	46
4	113	713
5	79	626

As we can see, sometimes we get extreme values for the number of 3-, 4- and 5-cycles. This can be explained, since a few times we should run into cases where some of the low-frequent symbols happen to form 8- or maybe 9-cycles. As we know from Theorem 1, one 8-cycle gives rise to eight 7-cycles. Each of the 7-cycles give rise to seven 6-cycles and so on, so the number of 4-cycles we get when there is an 8-cycle is $\binom{8}{4} = 70$. These numbers should contribute a lot to the average values in Table 2, but still the averages are only 16 for 3-cycles and 4.2 for 4-cycles. In the actual 340 cipher there are no cycles of length higher than 5, which makes the corresponding numbers 62 and 14 look even more extreme. Note that the number of 2-cycles in the cases where there is an 8-cycle is not that big, since $\binom{8}{2}$ is only 28. It should then be noted that in the 10.000 trials, never even once did we see as many as ninety 2-cycles. These facts indicate that the 340 cipher has not been constructed independently from

the cycle system, which again indicates that the 340 cipher is not just a random collection of ciphertext symbols. Thus we believe that there really is a plaintext message hidden there.

4. Solving the 408 cipher - Z408

In this section we give a short description of a software application we have made to "break" Z408. This section can be also viewed as a manual for it. The application is based on observations of Z408 discussed in the next section, and as far as we know it is the first application which can break Z408 in reasonable time (within minutes). The application together with Z408 translated into numeric form can be downloaded from [4] and it works in three phases:

(1) Finding n-cycles.

(2) Filtering partial key candidates.

(3) Final completion of words using a dictionary.

In the first phase we are looking for good n-cycles (also called groups). In the second phase we take possible combinations of candidates for n-cycles and filter bad partial plaintexts. In the third phase we try to complete partial plaintext using dictionary with 1000 of the most frequent English words.

4.1. Observations of Z408

Here we describe some observations of Z408, which were used to break Z408, and as we believe, also can be helpful to break Z340.

The cycle structure of Z480 has many errors [4], but there are no errors at the beginning of the ciphertext. There are no errors in cycles until the 129th symbol of the ciphertext.

Most of the ciphertext symbols appear in the first part of the ciphertext only once. There are two exceptions of this rule.

- Ciphertext symbols which are part of n-cycles ($n \geq 2$).
- Ciphertext symbols with small frequency forming 1-cycles (symbols which encode letters by 1–1 correspondence).

Since the first part of Z408 is encrypted by different symbols we can find whole words in the key. Table 4 shows the key of Z408.

The more frequent plaintext letters are enciphered by n-cycles for larger n. Table 5 shows cycle structure of Z408 with frequencies for each plaintext letter.

4.2. Finding n-cycles in Z408

As we have already mentioned earlier, each n-cycle can be built from 2-cycles (Theorem 1). Thus we start with finding 2-cycles first.

TABLE 4. Key of Z408.

ciphertext symbol	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
key	i	l	i	k	e	i	l	i	n	g	p	e	o	e	b	e	c	a
ciphertext symbol	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
key	u	s	e	t	s	s	o	m	h	f	n	t	s	o	r	e	f	n
ciphertext symbol	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54
key	t	h	a	n	w	d	a	e	t	o	r	r	a	d	v	l	y	x.

TABLE 5. Cycle structure of Z408.

Letter	n-cycle	freq
A	18 38 43 49	23
B	15	9
C	17	10
D	42 50	7
E	5 12 14 16 21 34 44	54
F	28 35	11
G	10	12
H	27 38	16
I	1 3 6 8	43
J		0
K	4	6
L	2 7 52	33
M	26	16

Letter	n-cycle	freq
N	9 29 36 40	23
O	13 25 32 46	28
P	11	7
Q		0
R	33 47 48	19
S	20 23 24 31	23
T	22 30 37 45	34
U	19	11
V	51	6
W	41	8
X	54	1
Y	53	8
Z		0

4.2.1. Finding 2-cycles in Z408

Due to errors in the cycle system of Z408 it is necessary to compute the probability $P_{(A,B)}$ that two given ciphertext symbols A, B form a cycle.

Let

$$S_{(A,B)} = (s_1, s_2, \ldots, s_k),$$

where

$$s_i \in \{A, B\} \quad \text{and} \quad s_1 = A,$$

denote the sequence of appearance of A, B ciphertext symbols throughout the whole cipher. A natural way of computing probability $P_{(A,B)}$ is to compare the number of alternating symbols in $S_{(A,B)}$ with a perfect 2-cycle of the same

length k. So we can compute $1 - P_{(A,B)}$ for A, B symbols as relative distance between a perfect 2-cycle and the sequence $S_{(A,B)}$. The distance between $S_{(A,B)}$ and a perfect 2-cycle of length k can be computed as

$$D_{(A,B)} = |AB + BA - (k-1)|,$$

where AB, BA represents numbers of pairs $s_i, s_{i+1} \in S_{(A,B)}$ such that

$$s_i = A,\ s_{i+1} = B \quad \text{or} \quad s_i = B,\ s_{i+1} = A.$$

The relative distance $DR_{(A,B)}$ between $S_{(A,B)}$ and a perfect 2-cycle of length k can be defined as the ratio $D_{(A,B)}/(k-1)$. This expression can be viewed as the probability of "not being a cycle" in the general case, and we used it for filtering of 2-cycle candidates.

Remark 1. In our application we assume only 2-cycles close to the perfect 2-cycle. In other words, the 2-cycle (A, B) is a candidate if its $P_{(A,B)}$ is greater than a threshold value.

Since we have knowledge about Z408 plaintext we can improve accuracy of finding real 2-cycles based on observations described in Section 4.1. From the observations we know that in the first part of the cipher there are no errors in the 2-cycles. So we can improve accuracy of finding a real 2-cycle by putting more weight to cycles appearing earlier. Also it is obvious that a longer uncorrupted 2-cycle form a real 2-cycle with higher probability. Based on previous facts we proposed a function, which computes rank for each pair of ciphertext symbols A, B. The pairs with high ranks will be used in the second stage to build n-cycles. The function uses the following parameters:

(1) lc – length of the longest uncorrupted cycle;
(2) n_1, n_2 – frequencies of symbols A, B in the ciphertext;
(3) bc – length of uncorrupted cycle from the beginning of the ciphertext.

We tested many weights for cycles, occurrence of symbols and the expression for rank. The best result has the following form:

$$\text{Rank} = 4lc + 2bc + n_1 + n_2.$$

Using this formula and taking 56 % as a threshold probability for $P_{(A,B)}$ we found the best 2-cycles summarized in Table 6. Each row of the table represents the 2-cycles with $P_{(A,B)} \geq 0.56$.

It should be noted that in the cipher Z408 only symbols $4, 10, 11, 15, 17, 19, 20, 41, 51, 53, 54$ encode letters by 1–1 correspondence, and thus do not form 2-cycles. This is in good agreement with Table 6. The only exception is 2-cycle $(17, 5)$. But as we can see the sequence $S_{(5,17)}$ do not pass the criteria which means that cycle $(17, 5)$ is close to the threshold.

TABLE 6. The 2-cycles of Z408 when $p_{(A,B)} \geq 0.56$.

A-symbol	B-symbols
1	3 6 8
2	7
3	6 8 1
5	12 14 16 21 31
6	8 1 3
7	2
8	1 3 6
9	29 21 36 13 18 31
12	14 16 21 31 5
13	25 9 32
14	16 21 31 5 12
16	21 5 12 14
17	5
18	9
21	5 12 14 16

A-symbol	B-symbols
22	30 37 45
25	46 13
27	38 45
29	31 36 9
30	37 45 22
31	5 12 14 9 29
32	13
33	47
34	47
36	5 12 14 9 29
37	45 22 30
38	22 30
45	22 30 37
46	25
47	34 33

4.2.2. Combining 2-cycles into n-cycles

In this step we choose the best candidates for 2-cycles, and combine them into n-cycles. An easy way to do this is to use the graph representation of the cycle structure of the cipher. Let $G = (V, E)$ be the graph representation of the cycle structure. The set of vertices V is the set of ciphertext symbols s_i. Vertices s_i, s_j are joined by an edge if and only if s_i, s_j form a 2-cycle. It is easy to see that the vertices of n-cycles form an n-clique. Therefore, to find n-cycles it suffices to use standard graph algorithms looking for cliques.

For n-cycles we can also compute a probability of being a cycle. Next we generalize the computation of $DR_{(A,B)}$ to n-cycles. Let $S_{(a_1,\ldots,a_n)}$ denote the sequence of appearance of the symbols $\{a_1, \ldots, a_n\}$ in the Z408 cipher. Let $O_{(a_{i_1},\ldots,a_{i_m})}$ represent the number of occurrences of the vector $(a_{i_1}, \ldots, a_{i_m})$ in the sequence $S_{(a_1,a_2,\ldots,a_n)}$. Then the distance between $S_{(a_1,a_2,\ldots,a_n)}$ and the perfect n-cycle of length k can be computed as $\operatorname{Max} O_{n,k} - \overline{O}_{(a_1,a_2,\ldots,a_n)}$, where $\operatorname{Max} O_{n,k}$ is the maximal possible value for $\overline{O}_{(a_1,a_2,\ldots,a_n)}$, and $\overline{O}_{(a_1,a_2,\ldots,a_n)}$ is

defined as

$$\begin{aligned}\overline{O}_{(a_1,a_2,\ldots,a_n)} &= O_{(a_1,a_2)} + O_{(a_2,a_3)} + \ldots O_{(a_n,a_1)} + \\ &\quad O_{(a_1,a_2,a_3)} + \ldots + O_{(a_n,a_1,a_2)} + \\ &\quad \vdots \\ &\quad O_{(a_1,a_2,\ldots,a_n)} + O_{(a_2,a_3,\ldots,a_n,a_1)} + \ldots + O_{(a_n,a_1,\ldots,a_{n-1})}.\end{aligned}$$

Remark 2. The maximum value $\operatorname{Max} O_{n,k}$ for an n-cycle of length k is given as $\operatorname{Max} O_{n,k} = (n-1)k - n(n-1)/2$.

The distance $D_{(a_1,a_2,\ldots,a_n)}$ is computed as

$$D_{(a_1,a_2,\ldots,a_n)} = \operatorname{Max} O_{n,k} - \overline{O}_{(a_1,a_2,\ldots,a_n)}$$

and the relative distance $DR_{(a_1,a_2,\ldots,a_n)}$ as

$$DR_{(a_1,a_2,\ldots,a_n)} = D_{(a_1,a_2,\ldots,a_n)} / \operatorname{Max} O_{n,k}.$$

Example 4.2.2 clarifies the computation of $DR_{(a_1,a_2,\ldots,a_n)}$.

EXAMPLE. We compute $DR_{(1,2,3,4)}$ where the sequence $S_{(1,2,3,4)}$ of length $k = 12$ has the following form

$$S_{(1,2,3,4)} = \{1,2,3,4,1,3,2,4,1,1,2,3\}.$$

Since we can find 2 vectors $(1,2)$ in the sequence $S_{(1,2,3,4)}$ then $O_{(1,2)} = 2$. Similarly we get

$$O_{(2,3)} = 2, \quad O_{(3,4)} = 1, \quad O_{(4,1)} = 2.$$

For vectors of the length 3 we get

$$O_{(1,2,3)} = 2, \quad O_{(2,3,4)} = 1, \quad O_{(3,4,1)} = 1, \quad O_{(4,1,2)} = 0.$$

Finally,

$$O_{(1,2,3,4)} = 1, \quad O_{(2,3,4,1)} = 1, \quad O_{(3,4,1,2)} = 0, \quad O_{(4,1,2,3)} = 0$$

and $\overline{O}_{(1,2,3,4)}$ is equal to the sum of all O-values, e.g., $\overline{O}_{(1,2,3,4)} = 13$. By Remark 2 we have

$$\operatorname{Max} O_{12,4} = 30, \quad D_{(1,2,3,4)} = 30 - 13 = 17, \quad DR_{(1,2,3,4)} = 0.57.$$

Therefore the given sequence has a 43 % probability to be a 4-cycle.

Using this formula and taking 70 % as a lower bound for $P_{(a_1,\ldots,a_n)} = 1 - DR(a_1,\ldots,a_n)$ values we found the n-cycles shown below. These correspond with real n-cycles of Z408 (Table 5).

n-cycles found:

```
5 12 14 16 21
22 30 37 45
 1  3  6  8
 2  7
```

4.3. Filtering partial key candidates

In this phase we try to filter out bad partial plaintexts.

First we find a possible set of candidates for each n-cycle. This was made by comparing ordinary frequencies of English letters with sum of frequencies of n-cycle symbols in the ciphertext. Assuming that a given n-cycle is a part of bigger cycle, it is necessary to take as candidates for n-cycles letters with higher ordinary English frequency. Of course, we have to assume some differences between Z408 letter frequency and ordinary frequency of English letters. After candidates for each n-cycle were found we made by exhaustive search partial keys and corresponding partial plaintexts. These partial texts were subsequently filtered by the following three filters:

- The first filter checks occurrence of impossible strings not found in the English language (like "AAAA", "SSSS", ...) formed by vowels and consonants. A partial key passes this filter if and only if there are no impossible strings in the partial plaintext.
- The second filter is based on the ability to cover partial plaintext using words from dictionary. Partial plaintext passes this filter if and only if for each plaintext letter there is a word which fits into the plaintext and cover that letter. Since there are usually few plaintext letters it is practical to apply this filter only on dense parts of the plaintext. In our application a user must choose the length of this part. The same number is used in the third filter.
- The third filter is based on the statistical length of English words. It is similar to second filter but it computes the average length of used words. A partial key passes this filter iff the average length of used words is approximately 3.79 which is the average length of words in all known Zodiac letters.

We performed several experiments to determine the strength of our filters. Table 7 summarizes the results of experiments.

Table 7. Experiments with filters.

	Exp1	Exp2	Exp3	Exp4
Keys - before Filter 1	1050	7260	96600	343200
Keys left after Filter 1	376	4561	60125	261265
Keys left after Filter 2	297	324	39474	195229
Keys left after Filter 3	14	36	4474	6088
Coverage in %	30	18.4	26.4	26.4
Time in seconds	1	15	783	1420

4.4. Completion words using dictionary

In this phase we take partially deciphered text, and try to complete words using a dictionary. This phase can be divided into two steps. First we try to expand actual partial key using some of the methods described below.

- The first method is based on average length of words which fit somewhere in the partial plaintext. We follow the conjecture that for correct plaintext and its parts, the average length of words is the biggest one. In this method we are looking for complete covering of the densest part of the plaintext with such property.
- The second method is similar to the first one but we take the densest part of the plaintext which starts at the first position.

Using the first and second method we get following partial keys and plaintexts.

```
FIRST METHOD
KEY:ilikeilingpeoebecauset*******t******t*******t*********
PARTIAL PLAINTEXT:
ilikekillingpeoplebecauseiti****uc**u*iti******u*t***killing*il*g
***i*t*e****estbec*u*e*a*i*t*e*o*t***g***u**n**al***llt*kill*o**t
*i*ggi*e**et*e***tt**i*li*ge*pe***c*itise*enbette*t***getti*g**u*
**ck*o***it**gi*lt**b**tpa*t**iti*t*****e*i*iei*i*lbe*eb**ni*p***
*ice***allt**i****kille**i*lbeco*e**s***e*i*illn*tgi*e**u*****ebe
cau****u*il*t**t*slo**o*****t*p**c***ectingo*sla*e********te*li*e
eb*o*iete**t**puti
COVERAGE: 54,25%

Second METHOD
KEY: ilikeilingpeoe*e****et*******t******t*******t*********
PARTIAL PLAINTEXT:
ilikekillingpeople*e****eiti**********iti*********t***killing*il*g
***i*t*e*****e*t*e*****e***i*t*e*o*t***g******n***l***llt*kill*o**t
*i*ggi*e**et*e***tt**i*li*ge*pe*****iti*e*en*ette*t***getti*g****
***k*o***it**gi*lt*****tp**t**iti*t*****e*i*iei*i*l*e*e***ni*p***
*i*e****llt**i****kille**i*l*e*o*e******e*i*illn*tgi*e*********e*e
*********il*t**t**lo**o*****t*p******e*tingo**l**e*********te*li*e
e**o*iete**t**p*ti
COVERAGE: 43,5%
```

In the next step we try to complete the most covered words. We found in the plaintext the words where at least 2/3 is covered and completed them. Subsequently we fill in new letters into the plaintext and key and repeat the process

again until no such words are left. In this step we are using only words of lengths $5, 6, 7$. After the completion of words we finally get following correct plaintexts.

```
KEY:ilikeilingpeoebecausetss*mhfnt****fnt******etorra**l**
RESULTING PARTIAL PLAINTEXT:
ilikekillingpeoplebecauseitiss*muchfuniti*m***funt***killing*il*g
*meintheforrestbecausemanist*emo*t***g***ue*namal*f*lltokillsomet
hinggi*esmet*em**tthrillinge*per**ceitise*enbette*t**ngetting**ur
rocksoff*ithagi*lt**bestpart*fiti*th****e*i*iei*illbereb**ninp*r*
*ice*n*allth*i***ekille**illbecomem*sla*esi*illn*tgi*e*oum*n*mebe
cause**u*illt**t*slo**o*nor*t*pm*collectingofsla*esf**m*afterlife
ebeorietemeth*puti
COVERAGE: 78,5%
```

```
KEY: ilikeilingpeoe*ec**set****hf*t****fnt*******torra**l**
RESULTING PARTIAL PLAINTEXT:
ilikekillingpeople*ec**seiti*****chf**iti*****f*nt***killing*il*g
***i*theforrest*eca**e**ni*t*e*o*t***g******na**l*f*lltokill*o**t
hi*ggi*e**et*e***tthrillinge*per**c*itise*en*ette*t***getting***r
rock*off*ithagi*lt******tp*rt*fiti*th****e*i*iei*ill*ere***ni*p*r*
*ice*n**llth*i****kille**ill*eco*e**sla*e*i*illn*tgi*e*o******e*e
c********illt**t*slo**o*nor*t*p**collectingofsl**e*f****afterlife
e**oriete**th*p*ti
COVERAGE: 61,25%
```

The plaintext has now appeared enough for the rest of the ciphertext to be solvable by hand.

5. Conclusion

In this paper we have described an idea for how one might proceed to solve the Z340 cipher. We have demonstrated strong evidence suggesting the cycle system evidently used in the solved Z408 also has been followed in the Z340. Next, we have shown how to look for the most probable cycles when there are errors in the cycle system.

Using this method we have been able to (re)solve the Z408. So far we have not been successful when attacking the unsolved Z340, but we believe it can be done using the methods described here if the parameters for the technique are set to appropriate values.

REFERENCES

[1] http://oranchak.com/zodiac/webtoy/
[2] http://www.ciphertool.com/
[3] http://www.zodiackiller.com/340Cipher.html
[4] http://sjgt.yweb.sk/tim/index.htm

Received April 30, 2010

Håvard Raddum
Department of Informatics
University of Bergen
N–5020 Bergen
NORWAY
E-mail: havard.raddum@ii.uib.no

Marek Sýs
Department of Applied Informatics
Information Technology
Slovak University of Technology
Ilkovičova 3
SK–812-19 Bratislava
SLOVAKIA
E-mail: marek.sys.@stuba.sk