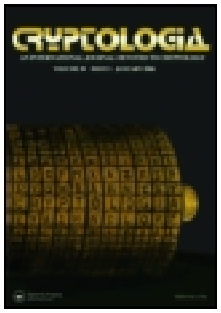

Cryptologia

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ucry20

STATISTICAL ANALYSIS OF THE HAGELIN CRYPTOGRAPH

Ronald L. Rivest

Available online: 04 Jun 2010

To cite this article: Ronald L. Rivest (1981): STATISTICAL ANALYSIS OF THE HAGELIN CRYPTOGRAPH, Cryptologia, 5:1, 27-32

To link to this article: http://dx.doi.org/10.1080/0161-118191855751

STATISTICAL ANALYSIS OF THE HAGELIN CRYPTOGRAPH

Ronald L. Rivest

We derive here a formula which estimates how much ciphertext is needed to solve a cryptogram produced by a Hagelin cryptograph, using the cryptanalytic technique presented by Barker [1]. We shall see that no more than 8000 characters of ciphertext are needed to solve a Hagelin Model C-48 (or U.S. Army M-209) cryptogram. The Hagelin cryptograph was invented in the 1930's by Boris Hagelin; many thousands of these machines were produced in the subsequent decades.

I. The Encryption Process

We let the letters of the alphabet used (both for plaintext and ciphertext) be denoted $a_1, a_2, \ldots, a_\lambda$, where λ is the number of letters (typically $\lambda = 26$).

A typical Hagelin machine has w keywheels, or wheels, where wheel i has t_i pins. For the C-48 we have w = 6 wheels of 17, 19, 21, 23, 25, and 26 pins, respectively. Each pin can either be pushed left or right. Each wheel rotates past a sensor; at a given moment the w sensors can determine which of the w wheels have "left"-pins under the sensor, and which have "right"-pins. Therefore one of 2^w possible sensor readings will occur; this sensor reading is used to select a monoalphabetic substitution to encrypt the plaintext letter. Each wheel then advances one position before the encryption of the next letter.

The detailed operation of the machine is described in Barker [1]. A substitution selected by a given sensor reading may be the same as substitutions selected by other sensor readings; this will not affect our analysis. (For example, on the C-48 we have $2^w = 64$ different readings possible but only 26 monoalphabetic substitutions available, so some substitutions will be selected by more than one reading.)

II. The Cryptanalytic Procedure

We assume that we must make a "ciphertext only" attack; no partial plaintext or probable words are available to us. We assume, however, that the plaintext is English (or another nonrandom source of characters). For English military text we can expect frequency counts per 1000 characters as shown in Table 1 (from [1], p. 109). The frequency of "Z" is high because the Hagelin machines are not equipped with a "space character, so Z is conventionally used in its place.

A-62	H-28	O-63	V-13
B-8	I-62	P-22	W-13
C-26	J-1	Q-3	X-4
D-35	K-2	R-64	Y-16
E-109	L-31	S-51	Z-162
F-24	M-21	T-77	
G-14	N-67	U-22	

Table 1. English Military Text Frequencies (per 1000)

The fact that the plaintext has a non-uniform distribution causes the ciphertext also to have a non-uniform distribution; this enables us to determine the pin settings using statistical tests on the ciphertext.

Suppose we wish to determine the pin-settings on wheel 1 (the other wheels can be handled similarly). Although a very large number of characters must be enciphered before the sequence of sensor readings repeats itself (this number is the least common multiple of $t_1,\ldots,t_w$), the pins passing under wheel 1's sensor will repeat every t_1 letters. If $C_1, C_2, \ldots$ is the sequence of ciphertext letters available, then pin 1 was used to encipher letters C_1, C_{t_1+1}, $C_{2t_1+1}, \ldots$ while pin 2 was used to encipher letters C_2, C_{t_1+2}, $C_{2t_1+2}, \ldots$, etc.

We wish to determine if pin i on wheel 1 has been pushed in the same direction as pin j on wheel 1, for all i and j, $i \neq j$. To do this we can compare the frequencies of the ciphertext letters produced using pin i with those produced using pin j. For example to compare pins 1 and 2 we can make a table such as is given in Table 2.

	a_1	a_2	...	a_λ	←Ciphertext Letter
pin 1	f_1	f_2	...	f_λ	n
pin 2	g_1	g_2	...	g_λ	n
total	f_1+g_1	f_2+g_2	...	$f_\lambda+g_\lambda$	2n

Table 2. Sample frequency table

Here f_i is the number of a_i's in C_1, C_{t_1+1}, $C_{2t_1+1}, \ldots$, while g_i is the number of a_i's in C_2, $C_{t_1+2}, \ldots$, etc. We have assumed for simplicity that the available ciphertext has length $t_1 \cdot n$, so that each pin of wheel 1 is used exactly n times.

If pins 1 and 2 are in the same position (i.e. either both left or right), we would expect the f_i's and g_i's to have the same distribution. On the other hand, if the pins are in different positions, the underlying non-uniformity of the plaintest distribution will result in a statistically significant difference in the distribution of the f_i's and the g_i's.

The X^2 test [2, p.447] will detect significant differences between the f_i's and the g_i's. We compute

$$X^2 = \sum_{i=1}^{\lambda} (f_i-g_i)^2/(f_i+g_i);$$

This statistic will follow the X^2 distribution with λ-1 degrees of freedom if the f_i's and the g_i's have the same distribution. Thus [2, p.234]

$$E(X^2) = \lambda-1, \text{ and} \tag{2}$$

$$\mathrm{Var}(X^2) = 2(\lambda-1). \tag{3}$$

If the f_i's and the g_i's are <u>not</u> from the same distribution, we can expect, that $X^2 \to \infty$ as $n \to \infty$.

When

$$X^2 > E(X^2) + 2\sqrt{Var(X^2)} = \lambda-1 + \sqrt{8(\lambda-1)} \qquad (4)$$

we may assume that the deviation is statistically significant.

Once we have enough ciphertext we will be able to determine which pins on wheel 1 are set in the same manner; that is we will have divided the pin positions into two groups, where every pin in a group is set in the same direction. Other simple techniques (see [1]) can then be used to decide which group is "left" and which "right". From there on determining the pin settings on the other wheels, etc., is relatively straightforward (see [1] for more details).

III. How Much Ciphertext is Needed?

Although the cryptanalytic procedure given above was published by Barker [1], no estimate was given there for the amount of ciphertext required to solve a cryptogram. The analysis given here uses only elementary techniques, and arrives at an answer which seems "reasonable" in comparison with the examples given in [1].

We first solve the following problem. Let p_i (respectively q_i) denote the probability of letter a_i occurring in the ciphertext when pin 1 (respectively pin 2) of wheel 1 is under the sensor. Then we want to know how much ciphertext is required to determine that the p_i's are different from the q_i's (assuming that they are different). The answer will of course depend on how different the two distributions are.

The distribution of f_i thus follows a binomial distribution with probability of success p_i:

$$\mathrm{Prob}(f_i=k) = \binom{n}{k} p_i^{\,k}(1-p_i)^{n-k} \qquad (5)$$

$$E(f_i) = np_i \qquad (6)$$

$$Var(f_i) = np_i(1-p_i). \qquad (7)$$

Similarly g_i follows a binomial distribution with probability q_i of success.

We assume that $p_i \approx q_i \approx 1/\lambda$ for all i in what follows. That is, we assume that each ciphertext letter will be (to a first-order approximation) equally likely; the X^2-test will however measure the second-order effects of any differences. Using this assumption in (1) to conclude that $f_i+g_i \approx 2n/\lambda$ we obtain

$$E(X^2) \approx (\lambda/2n) \sum_{i=1}^{\lambda} E((f_i-g_i)^2). \qquad (8)$$

Furthermore,

$$E((f_i-g_i)^2) = E(f_i^{\,2}) + E(g_i^{\,2}) - 2E(f_i)E(g_i) \qquad (9)$$

since f_i and g_i are independent random variables. Also

$$E(f_i^{\,2}) = Var(f_i) + (E(f_i))^2. \qquad (10)$$

Combining equations (6) - (10)

$$E(X^2) \approx (\lambda/2n) \sum_{i=1}^{\lambda} E(n\cdot(p_i(1-p_i)+q_i(1-q_i))+n^2(p_i-q_i)^2). \qquad (11)$$

Using our assumption that $p_i \approx q_i \approx 1/\lambda$ to simplify the coefficient of n in (11) we obtain

$$E(X^2) \approx (\lambda-1) + (\lambda n/2)\sum_{i=1}^{\lambda} E((p_i-q_i)^2) \qquad (12)$$

Note that this agrees with (2) when $p_i=q_i$ for all i; we obtain $E(X^2) = \lambda-1$ as we should. When the two distributions are different the excess of X^2 over λ increases linearly with n and with the sum of the squares of the differences in the corresponding probabilities. This completes our answer to our first problem for given probabilities p_i, q_i, since we can now calculate how large n must be for (4) to hold.

In order to use the above result we need to calculate $E((p_i-q_i)^2)$. Note that p_i and q_i are random variables; they depend on how the pins are set on wheels 2,...,w. Once we know the distribution of p_i and q_i we can calculate $E((p_i-q_i)^2)$, since

$$E((p_i-q_i)^2) = \mathrm{Var}(p_i-q_i) = 2\mathrm{Var}(p_i). \qquad (13)$$

This follows since $E(p_i) = E(q_i) = 1/\lambda$, so that $E(p_i-q_i) = 0$, and since $\mathrm{Var}(p_i-q_i) = \mathrm{Var}(p_i) + \mathrm{Var}(q_i)$, where $\mathrm{Var}(p_i) = \mathrm{Var}(q_i)$ since p_i and q_i will have the same distribution.

How is p_i determined? With pin 1 of wheel 1 in a fixed position, the other w-1 wheels can produce 2^{w-1} distinct sensor readings, each of which selects some monoalphabetic substitution. Since the recommended usage of the Hagelin machine is to set about half of the pins on each wheel in each direction, each of the 2^{w-1} substitution functions is equally likely to be used.

Let $\pi_1,\pi_2,\ldots,\pi_\lambda$ denote the respective probabilities of occurrence of $a_1,a_2,\ldots,a_\lambda$ in the plaintext. (For English military text these can be obtained from Table 1.) To determine the probability p_i of letter a_i occurring in the ciphertext when pin 1 is used, let $a_{j_1},a_{j_2},\ldots,a_{j_{2^{w-1}}}$ be the list of 2^{w-1} plaintext letters, the k-th of which causes a_i to be produced as the ciphertext when substitution k is used. (There may be repetitions in this list.) Then

$$p_i = (1/2^{w-1})\sum_{k=1}^{2^{w-1}} \pi_{j_k}; \qquad (14)$$

p_i is the mean probability of the 2^{w-1} plaintext letters which can produce a_i.

Let π be a random variable which is equally likely to take any one of the values $\pi_1,\pi_2,\ldots,\pi_\lambda$. Then

$$E(\pi) = 1/\lambda, \text{ and} \qquad (15)$$

$$\mathrm{Var}(\pi) = \sigma_0^2, \qquad (16)$$

for some value σ_0^2. From Table 1 we have $1/\lambda = .03846$ and $\sigma_0^2 \approx .001344$. Equation (14) says that p_i is the mean value of a sample of 2^{w-1} values of π, so that

$$E(p_i) = E(\pi) = 1/\lambda \qquad (17)$$

$$\mathrm{Var}(p_i) = \mathrm{Var}(\pi)/2^{w-1} = \sigma_0{}^2/2^{w-1} \tag{18}$$

Thus, although p_i and q_i both have expected value $1/\lambda$, the underlying non-uniformity of the distribution of the plaintext ($\mathrm{Var}(\pi) \neq 0$) will cause p_i and q_i to be samples from a distribution with non-zero variance. If pins 1 and 2 are set in the same manner, then $p_i = q_i$ is forced, but if they are set differently we may assume that p_i and q_i are independent, so that from (13) and (18) we obtain a non-zero $E((p_i-q_i)^2)$:

$$E((p_i-q_i)^2) = 2\mathrm{Var}(p_i) = \sigma_0{}^2/2^{w-1}. \tag{19}$$

Plugging this into (12) we obtain for the case that pins 1 and 2 are in different positions:

$$E(X^2) \approx (\lambda-1) + ((\lambda^2\sigma_0{}^2)/2^{w-1})\cdot n. \tag{20}$$

In order for this to be a significant deviation we want (4) to hold, so that

$$((\lambda^2\sigma_0{}^2)/2^{w-1})\cdot n \geq 2\mathrm{Var}(X^2) = \sqrt{8(\lambda-2)}. \tag{21}$$

Since $t_1 n$ is the total amount of ciphertext required to produce n ciphertext letters enciphered under each pin of wheel 1, we can rewrite (21) as

$$\text{Amount of ciphertext} \approx (t_1 2^w\sqrt{2(\lambda-1)})/(\lambda^2\sigma_0{}^2); \tag{22}$$

this amount of ciphertext should produce sufficient statistical evidence to determine all the pin settings.

Formula (22) is our main result. If we assume that we are trying to break a w-wheel Hagelin cryptogram where one of the wheels has length $t_1 = 17$, then we can calculate the amount of ciphertext required as estimated using (22) with $\lambda = 26$ and $\sigma_0{}^2 = .001344$:

Number of Wheels	Amount of Ciphertext
1	264 Characters
2	529 Characters
3	1058 Characters
4	2117 Characters
5	4233 Characters
6	8468 Characters

Table 3.

The amount of ciphertext required doubles with each additional wheel, reaching approximately 8000 characters for a six-wheel machine.

How realistic is this result? By way of comparison Barker [1] provides a set of four six-wheel problems to be solved. Since they all have the same pin settings, they can be combined for statistical purposes into a single problem of 3245 characters. Barker also solves as an example a 4-wheel problem of 770 letters in length.

We conclude that our analysis is probably a bit conservative; our estimates may be a factor of two to four to large. While we believe our analysis to be correct, several considerations may reduce the actual amount of ciphertext required:

(1) Our assumption that the X^2 statistic must be 2 standard deviations above its expected value in order to be considered significant is probably conservative; a cryptogram might still be easily breakable if the expected deviation in the case that the pins were set differently were only one standard deviation. This would reduce our estimates by a factor of two.

(2) There is a "snowballing" effect once several of the pin settings on a wheel have been correctly identified, since the statistics from the known settings can be combined to yield improved accuracy in the determination of the remaining settings.

(3) There may be characteristics of the Hagelin machines which we have ignored which permit more powerful statistical tests to be used. For example, it may be useful to know that if pins 1 and 2 are set in different positions, then the substitutions selected under pin 2 are all shifted by the same amount from the substitutions selected under pin 1 (this is the "lug setting" for wheel 1). In our analysis we assume that the substitutions were all randomly selected.

We leave as open problems the precise analysis of considerations (1) - (3). Since these considerations all tend to reduce the amount of ciphertext required, our estimate of 8000 characters to break a C-48 cryptogram should be taken as an upper bound. The reader is urged to devise improved statistical techniques and analyses which would provide an improved estimate of the amount of ciphertext required.

Acknowledgements
This research was supported by NSF grant MCS76-14294 and by the Xerox Corporation.

REFERENCES

1. Barker, W. G. 1977. *Cryptanalysis of the Hagelin Cryptograph.* Laguna Hill, CA: Aegean Park Press.

2. Cramer, H. 1974. *Mathematical Methods of Statistics.* Princeton: Princeton University Press.