Mathematical model for recovering discrete parts of a text message

Anastasia Malashina (amalashina@hse.ru)
HSE University https://orcid.org/0000-0002-5163-0593

Research Article

Keywords: Entropy, keyless method, stream ciphers

Posted Date: June 8th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-595013/v1

Mathematical model for recovering discrete parts of a text message

Anastasia Malashina

HSE University, Moscow, Russia

Abstract. The paper studies the procedure for restoring discreet segments of an unknown source message based on information about possible variants of each sign. An algorithm is proposed based on compiling dictionaries of appropriate lengths, searching for text sections with a total number of character variants that do not exceed a given boundary, and then iterating through and eliminating false variants of dictionary values. Statistical properties of short-length text dictionaries are investigated, and extrapolation estimates are made for long-length texts. The main mathematical properties of this algorithm are described. Theoretical studies of the effectiveness of the procedure under consideration are carried out within the framework of a certain probability-theoretical model.

Keywords: Entropy · keyless method · stream ciphers.

1 Introduction

In a number of tasks for the analysis of information security algorithms, a situation arises when information about possible variants of its characters appears in relation to the characters of an unknown source text message. This situation occurs, for example, when using stream encryption, if:

1) information about possible variants of the message characters was obtained through the side channels of the leak,

2) the power of the key set is less than the power of the plaintext alphabet,

3) the characters of the key sequence are not equally distributed,

4) the same encryption key is used repeatedly.

If the number of variants for each character is relatively small, then the original message is simply restored completely [1]. Otherwise, if the number of variants of signs can be considered random, it becomes possible to build a procedure for searching and then recovering discrete sections of the message on which signs are concentrated with a relatively small number of possible variants. This paper considers the recovery of messages in English with a fixed plaintext alphabet – 29 characters (26 letters of the Latin alphabet, a space, a period and a comma).

2 Description of the algorithm

Let's assume that for each character of the encrypted message, it was possible to construct a certain set of variants of its sign, among which is the true character. The number of such possible options may have a different probability distribution. In this paper, we consider the case of an equally probable distribution.

The algorithm consists of the following main steps:

1) Compilation of s-gram dictionaries. Dictionaries are compiled on the basis of a large text sample (text corpus).

2) Selection of "limited" segments of a message. Only those segments of a message that contain a relatively small number of possible variants are selected for recovery.

Selection criterion: the average geometric value of the number of possible options for the corresponding segment l_i does not exceed the specified critical boundary: $l_i = \sqrt[s]{l_{i_1} \cdot ... \cdot l_{i_s}} \leq L$, where s is the length of the segment, l_{i_j} is the number of character variants for the j-th character in the i-th s-gram.

3) Building options for recovering the segment. Iterate through all possible combinations that can be constructed using known character variants for each character of the selected segment.

4) Selection of all possible legal recovery options based on the dictionary. If the constructed version is a dictionary value, then it is considered that it represents a part of the legal text, and is considered as a potential segment of the original text of a message. If the compiled s-gram is not present in the dictionary, then it is assumed that it is a text of a random structure, and is rejected as a false recovery option.

5) Restore the message segment. Depending on the number of possible legal variants found, the message segment is considered successfully restored or not restored. The maximum allowed number of recovery options is estimated as: $k = 2^{\beta \cdot s}$, where s is the length of the message section, β is some constant < 1. In practice, $\beta = 0,1$ [6].

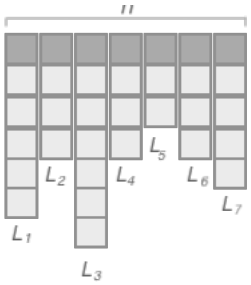

Fig. 1. Recovery Scheme

The following algorithm parameters are considered:

– Length of the text segment s: 10-25 characters
– Critical boundary of L: 8-16 characters
– $\beta = 0{,}1$

Text length	10	15	16	20	25
Number of possible options	1-2	1-2	1-3	1-4	1-5

Table 1. Acceptable degree of segment reconstruction ambiguity

3 Statistical properties of s-gram dictionaries

As part of the experimental study, dictionaries are compiled for segments of 10, 15, 20 and 25 characters long based on a text of 100 million characters in English.

Based on the data on the volume of the dictionaries obtained, an experimental assessment of the coverage of dictionaries and the entropy of the corresponding s-grams of the English language is carried out.

The compiled dictionaries were obtained on the basis of a limited volume of material and do not cover the set of all possible legal s-grams of the English language. Therefore, the coverage of the resulting dictionaries is evaluated. To estimate the coverage of the resulting dictionary, the number of s-grams that occur only once in the source material is used.

$$\tau = \frac{1 - n_s}{N_s}, \tag{1}$$

where N_s is the initial volume of the s-gram dictionary, n_s is the number of s-grams occurring once, τ is s-gram dictionary coverage.

Based on the amount of coverage, the volume of dictionaries is recalculated:

$$\tilde{N_s} = \frac{N_s}{1 - \frac{n_s}{N_s}}. \tag{2}$$

It is assumed that the value obtained in this way is the size of the real dictionary of s-grams of the language. Based on the theoretically found volume of the dictionary, the value of the entropy of s-grams is estimated as:

$$H_s = \log_2 \frac{N_s}{s} \tag{3}$$

Extrapolating these results to large values of s is difficult, since the shape of this sequence of values is unknown, except that it is positively decreasing.

The marginal entropy is defined as:

$$H_\infty = \lim_{s \to \infty} H_s \tag{4}$$

Dictionary Length	Volume	Entropy	Coverage (%)	Theor. dictionary	Entropy
10	6217191	2,29	19,95	∼31 million	2.49
15	9482897	1,57	7,21	∼131 million	1.80
20	10372296	1,17	3,27	∼317 million	1.41
25	10629589	0,93	2,07	∼513 million	1.16

Table 2. Dictionaries and entropy

To estimate the marginal entropy from this set of measurements, a model of sequential estimates is constructed.

It is assumed that the sequence of entropy values obeys a linear recurrence relation:

$$H_s - H_{s+1} = k \cdot (H_{s+1} - H_{s+2}) \tag{5}$$

The coefficient k for the model is determined numerically in accordance with the experimentally obtained entropy values for segments of small length.

In this case, the value of k, which gives the best approximation, is $k \approx 0.62$. By increasing the value of s, a sequence of heuristic estimates of H_s is constructed, the experimental evaluation of which becomes difficult for a large length of a text segment. Starting from $s = 50$, the values of H_s are stabilized and no longer change with the length of the segment.

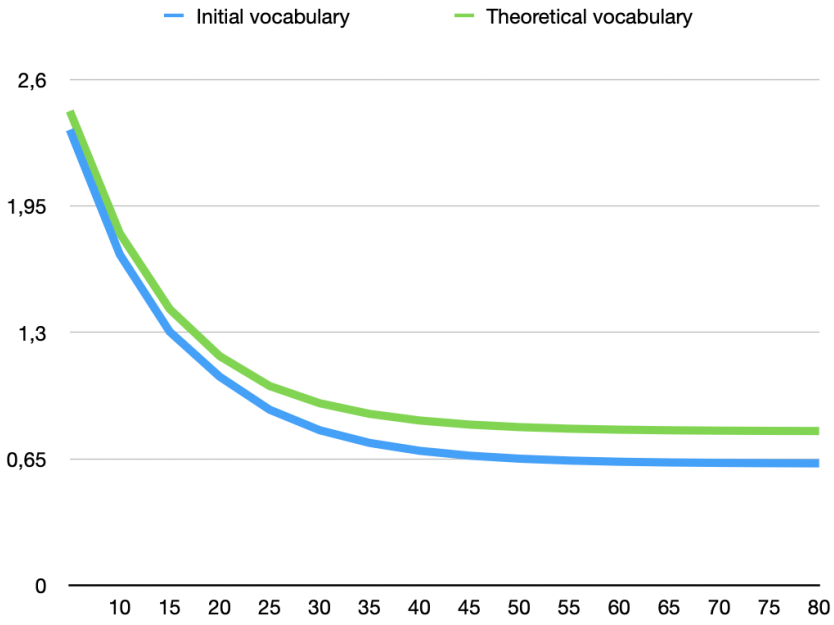

Fig. 2. Entropy

Thus, the maximum entropy is at $H = 0.8$ bits per character. The limit value can be used in theoretical calculations for large values of the message segment.

4 Mathematical properties of the algorithm

4.1 Probability of occurrence of bounded segments

Let the number of l_{i_j} sign variants for all segments be distributed independently and randomly equally probable from 1 to 29, that is, it takes values with the same probabilities $p_k = \frac{1}{29}$ for any $k = 1, 2, \ldots, 29$. The value of the critical boundary L is fixed.

Let's define the characteristic of the i-th segment of the message: $S_i = \frac{1}{s}\sum_{j=1}^{s} \log_2 l_{i_j}$.

In addition, for any segment, the expected value of S_i is $\mu = \frac{1}{29}\sum_{k=1}^{29} \log_2 k$ and the variance is $\sigma^2 = \frac{1}{29}\sum_{k=1}^{29} \log^2{}_L k - \left(\frac{1}{29}\sum_{k=1}^{29} \log_2 k\right)^2$.

Then the following theorems about the probability distribution of the appearance of a bounded segment in the message are valid.

Theorem 1. *Let the length of the message segment be $s \to \infty$. Then the probability that the geometric mean of the i-th segment does not exceed a set limit L: $\lim_{s\to\infty} P\left(\sqrt[s]{l_{i_1} \cdot l_{i_2} \cdot \ldots \cdot l_{i_s}} \le L\right) = \Phi\left(\frac{\log_2 L - \mu}{\sigma}\right)$ for any i, where Φ is a function of the standard normal distribution.*

Theorem 2. *Expected number of bounded segments of length s in a message of N characters: $(N - s + 1) \cdot P\left(\sqrt[s]{l_{i_1} \cdot l_{i_2} \cdots l_{i_s}} \le L\right)$.*

Theorem 3. *a) The correlation coefficient of two adjacent segments S_i and S_{i+1} with the length of s characters is equal to: $\rho = \frac{s}{s-1}$.*

b) The correlation coefficient of two arbitrary segments S_k and S_z with the length of s characters is equal to:

$$\begin{cases} \rho = \frac{s-z+k}{s}, & \text{если } z - k < s \\ \rho = 0, & \text{если } z - k \ge s \end{cases}$$

Theorem 4. *a) Conditional probability of occurrence of a bounded segment, provided that the previous segment is bounded:*

$\lim_{s \to \infty} P\left(S_i < \log_2 L \mid S_{i-1} < \log_2 L\right) = \frac{\int_0^{\log_2 L}\int_0^{\log_2 L} f(x,y)dxdy}{\Phi\left(\frac{\log_2 L - \mu}{\sigma}\right)}$,

where $f(x,y) = \frac{1}{2\pi\sigma^2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu)^2}{\sigma^2} - \rho\frac{2(x-\mu)(y-\mu)}{\sigma^2} + \frac{(y-\mu)^2}{\sigma^2}\right)}$, $\rho = \frac{s-1}{s}$ *is correlation coefficient S_i and S_{i-1}.*

b) The expected geometric mean value of the next segment, provided that the previous one is limited: $\lim_{s \to \infty} E\left(S_i \mid S_{i-1} < \log_2 L\right) = \mu - \frac{s-1}{s}\sigma\frac{\varphi\left(\frac{\log_2 L - \mu}{\sigma}\right)}{\Phi\left(\frac{\log_2 L - \mu}{\sigma}\right)}$, *where φ is the density of the standard normal distribution, Φ is a function of the standard normal distribution.*

Theorem 5. *The probability that all segments of the message are simultaneously bounded tends to a multidimensional normal distribution.*

$\lim_{s \to \infty} P\left(S_1 < \log_2 L,\ \ldots,\ \ S_{N-s+1} < \log_2 L\right) = N\left(\vec{\theta}, \Sigma\mathrm{s}\right)$,

where $\theta = (\mu, \mu, \cdots, \mu)$ *is a vector of average values,*

$\Sigma\mathrm{s} = \begin{bmatrix} \sigma^2 & \cdots & \sigma_{S_1,\ S_{N-s+1}} \\ \vdots & \ddots & \vdots \\ \sigma_{S_{N-s+1},\ S_1} & \cdots & \sigma^2 \end{bmatrix}$ *is a covariance matrix, at the intersection of the k-th line and the z-th column are the values of the covariance of the segments:* $\sigma_{S_k,\ S_z} = \mathrm{Cov}\left(S_k,\ S_z\right) = \frac{s-z+k}{s}\sigma^2$ *if* $z - k < s$. *Otherwise 0.*

Remark. The sum of independent uniform quantities quickly converges to the normal distribution. Traditional estimates of the error generated by the CLT in the final case, such as the Berry-Essen inequality, at an average value of s (tens) show a very rough estimate and an overestimated upper bound. This is due to the fact that, first, different classes of distributions converge to the normal at different rates, and the error estimate is given in general for any distribution structure. Secondly, the Berry-Essen inequality uses an error estimate in the form of the ratio of the third moment to the root of the number of terms. This form of error will give a fairly accurate estimate only on large s. It is known from numerical calculations that the uniform distribution converges quickly to the normal one, much faster than the general theoretical estimates of [5]. Thus, even for small values of s, the approximation by the normal distribution allows us to obtain a fairly accurate estimate of the theoretical probability.

4.2 Mathematical model of the distribution of the number of possible legal texts

Let $N = m^s$ be the total number of s–grams that can be constructed from characters of the alphabet with power m (in this paper $m = 29$ is fixed), D be the number of possible legal texts of length s in the same alphabet (this value is estimated as 2^{Hs}, where H is the entropy of s-grams), $n_i = l_i^s$ is the number of possible recovery options for the i-th segment of the message, k_i is the number of possible legal recovery options for the i–th recoverable message segment among n_i, where $l_i = \sqrt[s]{l_{i_1} \cdot \ldots \cdot l_{i_s}}$ – the average number of character variants for the unknown character of the i-th section of the message. In this case, it is assumed that the true recovery option is always present in the sample, that is, $n_i - 1$ are false options.

Then the probability that in a sample of n_i different options for the recovery obtained from the set of all possible variants of the N containing D legal variants, exactly $k_i$variants are legal, is described by the hypergeometric distribution [6], that is: $P\left(k_i\right) = \frac{C_D^{k_i} C_{N-D}^{n_i - k_i}}{C_N^{n_i}}$.

Theorem 6. *If a section of a message in English (alphabet power is 29 characters) with a length of s characters with an average number of variants of char-*

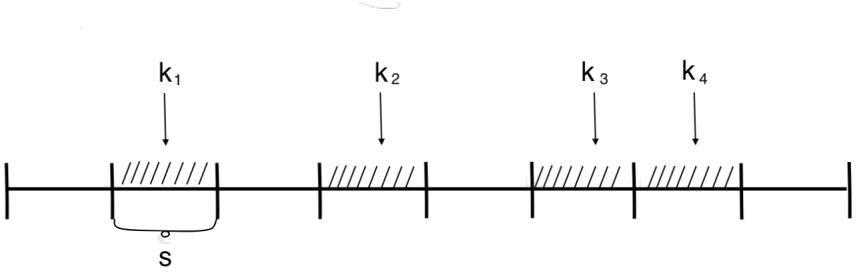

Fig. 3. Number of possible legal texts per segment

acters l_i is restored, then the probability of k_i possible legal variants of restoring this segment is: $P(k_i) = \frac{2^{H_s \cdot s}! \cdot \left(29^s - 2^{H_s \cdot s}\right)! \cdot l_i^s! \cdot (29^s - l_i^s)!}{k_i! \cdot (2^{H_s \cdot s} - k_i)! \cdot \left(l_i^s - k_i\right)! \cdot \left(29^s - 2^{H_s \cdot s} - l_i^s + k_i\right)! \cdot 29^s!}$.

The most likely number of possible legal texts (taking into account the true one) that will be found when restoring the message segment (hypergeometric distribution mode):

l_i/s	**10**	**15**	**20**	**25**
8	3	1	1	1
10	24	1	1	1
12	143	2	1	1
14	667	11	1	1
16	2534	80	1	1

Table 3. Most likely number of possible legal texts

Therefore, restoring 10-grams with an average number of character variants greater than 8-9 characters is practically meaningless. For 15-grams, the maximum recovery efficiency for L is no more than 12-13 characters.

As the length of the message segment increases, the degree of ambiguity in restoring the plaintext decreases. However, for large sections of text, the process of compiling search dictionaries becomes very difficult. The simulation allows us to choose the most **optimal algorithm parameter** is the length of the recovered message segment-based on the most effective ratio between the degree of ambiguity of the plaintext recovery and the coverage of the corresponding dictionary. Obviously, such a parameter is the minimum length of the restored s-gram, at which the probability of restoring the segment tends to one for any average number of true sign variants. Based on the probabilities of the hypergeometric distribution, the length of such a minimal s-gram is determined **16 characters**.

The limit distribution of the number of possible legal texts occurs when $s \to \infty$. Then the parameters of the hypergeometric distribution are $N = 29^s \to$

$\infty, D = 2^{H \cdot s} \to \infty, n = l_i^s \to \infty,\ l_i \leq 21, H = 0.8$. The type of marginal distribution depends on the number of possible legal texts in the sample.

Theorem 7. *The probability of finding exactly 1 legal text (true) at* $s \to \infty$*:* $P(1) \approx e^{-2\left(\frac{l_i \cdot 2^H}{29}\right)^s}$

Theorem 8. *If the number of possible legal texts is* $k = 2^{s \cdot \beta}$*, then the probability of getting* k *possible legal texts when restoring a segment of length* $s \to \infty$ *is:* $P\left(k = 2^{\beta s}\right) \approx \frac{1}{\sqrt{\pi}} 2^{((H-\beta+\log_2 \frac{l_i}{29})s + \log_2 e) \cdot 2^{\beta s} - \frac{\beta}{2}s - \frac{1}{2}}$.

5 Algorithm efficiency

The probability of restoring a bounded segment of a given length is the joint probability of such a segment appearing in the message and the success of its restoration, taking into account the acceptable polysemy:

$$P_{recovery}(s, l_i) = P_{occurrence}(s, l_i) \cdot P_{uniqueness}(s, l_i), \tag{6}$$

where $P_{occurrence}(s, l_i)$ is the probability of occurrence of a segment of length s and with the average geometric sign variants l_i, $P_{uniqueness}(s, l_i)$ is the probability that the segment recovery does not exceed the allowed polysemicity, $P_{recovery}(s, l_i)$ is the probability of successful recovery of the segment.

Then the probability of restoring a message segment whose average value of the character variants does not exceed the specified limit is estimated as: $\sum_{l_i < L} P_{recovery}(s, l_i)$.

Theoretical evaluation of the algorithm efficiency (the proportion of recovered segments in the message) for different parameter values:

L	10	15	16*	20	25
< 8	0,014	0,006	0,005	0,002	0,001
< 10	0,016	0,065	0,060	0,041	0,027
< 12	0,016	0,213	0,243	0,219	0,193
< 14	0,016	0,256	0,512	0,514	0,515
< 16	0,016	0,256	0,745	0,769	0,795
< 18	0,016	0,256	0,887	0,912	0,929
< 20	0,016	0,256	0,956	0,972	0,984
< 22	0,016	0,256	0,984	0,992	0,996
< 24	0,016	0,256	0,995	0,998	0,999

Table 4. Theoretical fraction of recovered segments

Thus, with an increase in the average number of character variants of an unknown text segment with a length of at least 16 characters, the probability of its recovery increases, taking into account the permissible degree of ambiguity.

6 Conclusion

In this paper, we propose an algorithm for restoring individual sections of encrypted messages, which can be used in a number of cases, such as the use of an incomplete or uneven key sequence, leakage through side communication channels, and repeated use of the same encryption key. The recovery procedure is based on compiling dictionaries of the appropriate lengths, selecting message segments that contain characters with a relatively small number of possible variants, sorting through all possible combinations, and eliminating false variants that are not dictionary units. The theoretical evaluation of the proposed method is carried out in the framework of an equally probable model and the optimal parameters of the algorithm are found. As part of the experimental study, s-gram dictionaries were created and their statistical properties were studied. A method of theoretical evaluation of dictionary coverage is developed. The values of the entropy of s-grams are found and the limit value of entropy is estimated. Various probability distributions arising in the framework of this algorithm are considered. It is found that as the length of the segment to be restored increases, the probability of finding a bounded segment approaches the normal distribution. Limit distributions are given for the probability of occurrence of a single segment, for the conditional probability of occurrence of a bounded segment, and for the joint occurrence of bounded segments. It is found that the normal distribution allows us to obtain a good approximation for small values of the segment length (tens). In addition, it is shown that the degree of ambiguity arising in the considered problem of restoring text segments is described by a hypergeometric distribution, and some numerical calculations of the probability of the appearance of an acceptable number of possible legal texts during the restoration are given. The asymptotics of the hypergeometric distribution for large values of the segment length are found. Using the found probability distributions, the general theoretical efficiency of the algorithm under consideration is estimated – the average fraction of the recovered segments. In the future, it is planned to consider other probabilistic distributions of the number of variants of the message symbol (polynomial). Investigate the probability distribution that occurs when the key is reused, and evaluate the effectiveness of the algorithm in this case.

References

1. Alferov, A. P., et al. "Foundations of cryptography." Gelios ARB, Moscow (2002)
2. Babash A. V. and Shankin G. P., "Cryptography." Solon-Press, Moscow (2007)
3. Bauer, Heinz. Measure and integration theory. Vol. 26. Walter de Gruyter, 2011.
4. Greene, William H. "Limited dependent variables—truncation, censoring and sample selection." Econometric analysis (2008): 833-902.
5. Zolotarev, Vladimir M. Modern theory of summation of random variables. Walter de Gruyter, 2011.
6. Proskurin G. V. Principles and methods of information protection, MIEM, Moscow (1997)