ISSN 1064-5624, Doklady Mathematics, 2023, Vol. 108, Suppl. 2, pp. S282–S292. © Pleiades Publishing, Ltd., 2023.
Russian Text © The Author(s), 2023, published in Doklady Rossiiskoi Akademii Nauk. Matematika, Informatika, Protsessy Upravleniya, 2023, Vol. 514, No. 2, pp. 138–149.

Possibility of Recovering Message Segments Based on Side Information about Original Characters

A. G. Malashina[a],*

Presented by Academician A.L. Semenov

Received September 4, 2023; revised September 15, 2023; accepted October 24, 2023

Abstract—To provide secure information exchange in communication channels, the correctness of the operation of the relevant information protection systems must be preliminary studied. The mathematical algorithms used in such systems are correct and can theoretically provide the correct statistical properties of the output stream compared to the input. However, at the stage of implementation (programming) of these protection algorithms or at the stages of assembling the final equipment (using hardware, making adjustments) and its operation in real conditions, it is possible to introduce distortions that violate the operation of certain elements of information security tools (for example, a random number generator). As a result, by the nature of the transmitted signal, it becomes possible to determine that the output stream for a number of characteristics is steadily different from the ideal encrypted stream, which in theory should have come from the equipment and appeared at the output of the communication channel. In this situation, it is necessary to understand how the introduction of certain distortions affects the degree of security of the system being created. For this purpose, the parameters of various message sources are described, which simulate receiving an output stream with distortions. At the same time, the degree of security of the corresponding communication channel is proposed to be determined by estimating the proportion of the input stream that can be restored from the output using side information resulting from the introduction of appropriate distortions in the operation of the system.

Keywords: legal texts, dictionary attack, communication channels, interception channels

DOI: 10.1134/S106456242370151X

1. INTRODUCTION

2.1. *Introduction to the Problem*

During the implementation and operation of some information security systems, it is possible to introduce disturbances into the operation of various elements of equipment (for example, a random number generator), distorting the correct result of applying mathematical encryption algorithms.

Such violations can lead to the cases listed below [1–3, 20, 21]:

- reducing cardinality of the key alphabet (when using a running key ciphering) or having unequal distribution to it;
- leaks of information about the characters of the original message;
- reuse of the encryption key.

Thus, as a result of the distorted operation of the mathematical encryption algorithm, side information about the characters of the input message appears at the output of the communication channel. Given this information and assuming meaningfulness[1] of the original message in the language in question, you can build an algorithm to restore the original text or its individual parts.

The full recovery is possible if the number of expected meanings of the original message characters is limited and the probability of a true character appearing among them is close to one. Then, the set of all its possible values is fixed for each input character. Recovering the source text is a search for a meaningful text among all possible combinations [4, 5].

This approach can lead to the loss of a true recovery variant, the probability of which is estimated based on the introduced probabilistic theoretical model of the process under study [6–8]. As the number of values at the output of a communication channel increases, restoration of the original text becomes difficult due to the high degree of uncertainty in choosing a meaningful text that arises in the process of searching for suit-

[a] *HSE University, Moscow, Russia*
**e-mail: amalashina@hse.ru*

[1] The meaningfulness of a text means the admissibility of its existence in the language.

able options. That is, the difficulty lies in selecting the true source text among a large number of recovered meaningful texts.

The degree of reconstruction ambiguity r is the mathematical expectation of the number of possible meaningful texts that can be obtained by applying the recovery algorithm [2]. The value of r depends on the length of the text s and allows evaluating the effectiveness of restoring the source text. The goal of the recovery procedure is to obtain an authentic, meaningful text.

Thus, when the number of characters at a channel's output is small, the meaningful text can be constructed in only one way, because all other combinations will turn out to be the text of a random structure. However, as the number of values increases, this approach results in more than one possible recovery option being found. In this case, it is impossible to determine which of the found texts is the original message without additional information. Therefore, using this procedure to recover a message becomes entirely ineffective.

However, the question arises about the possibility of recovering individual parts of unknown message. If the corresponding values of character sets are the implementation of a random variable, then, with a sufficient length of the message, rare favorable events may appear when the sets of small length appear in certain segments. On such segments it is possible to construct an algorithm for determining the required part of an unknown text, for example, by preliminary compiling the dictionary values of the appropriate length.

Problem Formulation

Suppose that we transmit the characters of the input message $x_1,\dots,x_N$, selected from the alphabet A of cardinality m, and obtain the set of characters of variable lengths $\{x_i\}_{l_i} = (x_i^{(1)},\dots,x_i^{(l_i)})$, $x_i^{(j)} \in A\ \forall i = 1, \dots, N$, $j = 1,\dots,l_i$, at the output of the communication channel. The values of l_i are assumed to be such that complete recovery of the message is impossible. In this case, the task is in searching for message sections with s characters with relatively small values of l_i and attempting to reconstruct individual parts of the transmitted message on them. Within the framework of a certain probability-theoretic model of the appearance of sets $\{x_i\}_{l_i}$ and dictionary models, it is required to estimate the average proportion of the recovered text of the original message.

2. MATHEMATICAL MODEL OF COMMUNICATION CHANNEL

Let us give a description of a communication channel that models reception of an output stream when disturbances are introduced into the operation of elements of information security systems.

Suppose that for each message character we can construct a certain set of values, among which the true desired character is present. The number of such possible options may have different probability distributions.

Let us consider the corresponding discrete communication channel without memory [9, 10]. Suppose that a set of characters A ($|A| = m < \infty$) acts as finite input and output alphabets. Each ith input character of the message x_i corresponds to the set $\{x_i\}_{l_i}$ with a cardinality l_i.

Let us set the probabilities $p_k^{(i)} = P(l_i = k)$ that the set $\{x_i\}_k$, consisting of k values; $k = 1,\dots,m$; $\sum_{k=1}^{m} p_k^{(i)} = 1$, appears at the output of the set.

Next, we determine that the composition of the set of values $\{x_i\}_k$ for each input character is formed by a random and equally probable selection from all ordered sequences of length k, constructed without repeating characters in the original alphabet.

Thus, the channel model is defined:

- by the alphabet A of code characters at the input and output with cardinality m;
- by the set of input messages X with N characters, where $x = (x_1, x_2,\dots,x_N) \in X$ is the input message, $x_i \in A$, $\forall i = 1,\dots,N$;
- by the set of values of output characters $(\{x_1\}_{l_1}, \{x_2\}_{l_2},\dots,\{x_N\}_{l_N})$, where l_i is the cardinality of the set $\{x_i\}_{l_i}$, $\forall i = 1,\dots,N$;
- by the probabilities $p_k^{(i)}$ of appearance of the output sets $\{x_i\}_k$ for $k = 1,\dots,m$, while the composition of the set $\{x_i\}_k$ is formed by selecting each character from the alphabet A according to the urn scheme without return, $\forall i = 1,\dots,N$.

Next, we propose an algorithm for reconstructing individual parts of an unknown message transmitted in a given communication channel and consider the case of a fixed arbitrary probability distribution $p_k^{(i)} = p_k$ in the communication channel for all $i = 1,\dots,N$.

3. DESCRIPTION OF THE ALGORITHM

3.1. s-Grams of Message

To implement the recovery algorithm, the message is divided into separate s-grams. Under s-gram we understood a sequence of s characters of text selected "with engagement," that is, for each subsequent s-gram there is a shift to the right by one character.

The number of s-grams in a message is determined by the serial number of its first character. That is, ith s-gram is denoted by $(x_i, x_{i+1}, \ldots, x_{i+s-1})$. When considering an individual s-gram, we introduce the internal numbering of its characters. For example, for the ith s-gram: $(x_{i_1}, x_{i_2}, \ldots, x_{i_s})$. The corresponding number of values for the jth character of the ith s-gram is denoted by l_{i_j}. The total number of s-grams per message with a length of N characters equals $N - s + 1$.

3.2. Algorithm for Recovering s-Grams

Algorithm input: the total length of unknown message (N), the length of restored sections (s), information about the possible values of each ith message character ($\{x_i\}_{l_i}$) and their quantity (l_i), the critical threshold of segment selection (L), the parameter (β), the prebuilt dictionary of s-grams and its cardinality (D_s).

The algorithm includes the following steps:

1. Selecting the appropriate message segments. For restoring, we need to select those s-grams of a message for which a small average number of values is known.

Criterion for selecting the s-gram: the geometric mean L_i of the number of pole characters for the corresponding ith s-grams must not exceed a given critical threshold L for all i:

$$L_i = \sqrt[s]{l_{i_1}, \ldots, l_{i_s}} < L,$$

where s is the length of the text segment (s-gram), l_{i_j} is the number of options for the jth character in the ith s-gram, L_i is the average number of options for the ith s-gram, and L is the critical threshold for segment selection.

2. Building the options of s-gram recovery. Finding all possible combinations that can be constructed using known information about the values of the characters in the s-gram.

3. Selecting the meaningful recovery options. If the composed text of the s-gram is present in the dictionary, then it is considered meaningful and is accepted as a possible recovery option of the true s-gram of the message. If no dictionary matches are found, this recovery option is rejected.

4. Recovering the s-gram of the message. According to the number of meaningful s-grams that we can construct, the selected segment of the message is considered restored or unrecovered.

Criterion for recovering s-grams: the degree of recovery uncertainty (the number of recovery options for one segment) should not exceed $r \leqslant r_{\max} = \lfloor 2^{\beta s} \rfloor$, where s is the length of the message segment and β is the numerical parameter less than 1. In practice, the value is often used $\beta = 0.1$ (Table 1).

Table 1. The admissible degree of ambiguity in reconstructing a segment, $\beta = 0.1$

Text length, s	10	15	16	20	25
Number of possible variants, $r_{\max}$	2	2	3	4	5

Algorithm output: the options for restoring individual s-grams of the message, the number of recovered s-grams.

The stage of creating short dictionaries of s-grams of a given coverage is not included in this algorithm. We preliminary created dictionaries on the basis of the corresponding text corpus [12, 13] (more details in Section 4).

For numerical evaluations, the following algorithm parameters are considered:

1. text segment length s: 10–25 characters;
2. critical threshold L: 8–16 characters;
3. parameter $\beta = 0.1$.

The main parameter of the algorithm we need to determine is the average share of recovered information at the output for a given probability distribution in the communication channel.

4. DICTIONARIES of s-GRAMS

Dictionaries are sets of s-grams arranged in alphabetical order without repetition. The process of creating a dictionary consists of extracting all s-grams from some language corpus and removing duplicate s-grams before sorting.

Since dictionaries are compiled from a limited language corpus, their coverage is incomplete. This means that not all existing s-grams of the language are included in this dictionary. That is, errors may occur when the existing s-gram is not in the dictionary and is discarded as invalid in the given language.

Thus, the degree of coverage τ_s of a dictionary of s-grams is the ratio of the size of the constructed dictionary to the total number of existing s-grams in the language [14]:

$$\tau_s = \frac{D_s}{M(s)},$$

where D_s is the size of the dictionary of s-grams and $M(s)$ is the number of meaningful s-grams in the language.

The issue with determining the degree of coverage is that the exact number of all meaningful s-grams in the language is unknown, especially for large orders, although an asymptotic estimate can be obtained using Shannon's probability-theoretic model [22]. To roughly estimate the degree of coverage for the dictionaries of s-grams, we can also use other methods: empirical verification [11, 14], estimation based on the number of one-time occurrences of s-gram [13], etc.

Dictionaries of s-grams are a plaintext model in which a message is treated as a sequence of s-grams from the dictionary. In the proposed algorithm, these dictionaries are considered a criterion for plaintext recognition.

Incomplete coverage may result in the loss of a true recovery option of s-gram if it is not in the dictionary we are using. Therefore, τ_s is a parameter that affects the overall probability of successfully recovering a message segment in the communication channel at step 4 of the algorithm.

Based on the size of the dictionary, we calculate the entropy of s-gram per character [12, 13]:

$$H_s = \frac{\log_2 D_s}{s}. \tag{1}$$

The entropy values of s-grams are used to estimate the number of meaningful texts of a given length in a language [15, 16].

The limit value H_s is taken as the entropy of a language:

$$H = \lim_{s\to\infty} H_s. \tag{2}$$

5. MATHEMATICAL PROPERTIES OF THE ALGORITHM

5.1. Probability of Appearance of L-Limited Segments

The number of possible values l_{i_j} for an unknown i_jth character is a discrete random variable taking integer values from 1 to m with probabilities $p_k = P(l_{i_j} = k)$ for any i_j. In this case, the random variables l_{i_j} for all segments are distributed equally and independently. The critical threshold value L is fixed, and the value of m is finite.

The message segment (s-gram) with number i is called L-limited if

$$L_i = \sqrt[s]{l_{i_1}\cdot l_{i_2}\cdot\ldots\cdot l_{i_s}} \le L. \tag{3}$$

The average characteristic of the ith message segment is called a random variable:

$$S_i = \frac{1}{s}\sum_{j=1}^{s}\log_2 l_{i_j}. \tag{4}$$

The probability of occurrence of a message segment selected for recovery is determined by the probability of its L-limitedness and can be expressed through the quantity S_i:

$$P_{\text{occur}}(s,L) = P\{\sqrt[s]{l_{i_1}\cdot l_{i_2}\cdot\ldots\cdot l_{i_s}} \le L\} = P\{S_i \le \log_2 L\}. \tag{5}$$

Because for any l_{i_j} it is true that $\mu = (\log_2 l_{i_j}) = \sum_{k=1}^{m} p_k \log_2 k$, $\sigma^2 = \mathrm{D}(\log_2 l_{i_j}) = \sum_{k=1}^{m} p_k \log_2^2 k - \left(\sum_{k=1}^{m} p_k \log_2 k\right)^2$, the mathematical expectation and the variance of the quantity S_i are equal to $\mathrm{E}S_i = \mu$ and $\mathrm{D}S_i = \frac{\sigma^2}{s}$ for any i.

Let the length of the message segment be $s \to \infty$. Then, by the central limit theorem [17],

$$\lim_{s\to\infty}\left(P\left(\sqrt{s}\cdot\frac{S_i-\mu}{\sigma}\right) - \Phi(R)\right) = 0$$

for any i, where $R = \sqrt{s}\cdot\frac{\log_2 L - \mu}{\sigma}$ and Φ is the function of standard normal distribution.

This limit distribution can be used to approximate the probability of occurrence of L-limited segments in the message when the quantity s is finite:

$$P_{\text{occur}}(s,L) \approx \Phi(R).$$

Because the message has a length of N characters, the number of consecutive segments of length s is $N - s + 1$. The expected number of L-limited segments in the message is

$$(N - s + 1)\cdot P_{\text{occur}}(s,L). \tag{6}$$

The values S_{i-1} and S_i of adjacent segments are dependent (have a nonempty intersection), because they have $s - 1$ common components. Let us determine the degree of their correlation dependence showing the proportion of matching elements in adjacent segments.

(a) The correlation coefficient of two adjacent segments of length s characters is

$$\rho = \frac{s-1}{s}. \tag{7}$$

(b) The correlation coefficient of the kth and mth ($k < m$) arbitrary segments with a length of s characters is

$$\rho = \begin{cases} \dfrac{s-m+k}{s} & \text{if} \quad m-k < s \\ 0, \quad \text{otherwise.} \end{cases} \tag{8}$$

(a) Note that $l_{i-1_j} = l_{i_{j-1}}$. Because the variables S_{i-1} and S_i have a nonempty intersection, the correlation coefficient between their average values reads

$$\rho = \frac{\mathrm{Cov}(S_{i-1}, S_i)}{\sqrt{\mathrm{D}S_{i-1}}\sqrt{\mathrm{D}S_i}} = \frac{\mathrm{Cov}\left(\log_2 l_{i-1_1}, \sum_{j=2}^{s}\log_2 l_{i_j}\right)}{s^2\sqrt{\mathrm{D}S_{i-1}}\sqrt{\mathrm{D}S_i}}$$
$$+ \frac{\mathrm{Cov}\left(\sum_{j=2}^{s}\log_2 l_{i_j}, \sum_{j=2}^{s}\log_2 l_{i_j}\right)}{s^2\sqrt{\mathrm{D}S_{i-1}}\sqrt{\mathrm{D}S_i}}$$
$$+ \frac{\mathrm{Cov}(\log_2 l_{i-1_1}, \log_2 l_{i-1_{s+1}})}{s^2\sqrt{\mathrm{D}S_{i-1}}\sqrt{\mathrm{D}S_i}}$$
$$+ \frac{\mathrm{Cov}\left(\sum_{j=2}^{s}\log_2 l_{i_j}, \log_2 l_{i-1_{s+1}}\right)}{s^2\sqrt{\mathrm{D}S_{i-1}}\sqrt{\mathrm{D}S_i}} = \frac{s-1}{s}.$$

(b) Note that S_k is a linear combination of random variables. This formula is generalized for two arbitrary segments in a similar way, taking into account the covariance property: if $m - k < s$, then S_k and S_m have a nonempty intersection and each of the two components has $m - k$ various independent terms, and the rest $s - m + k$ terms coincide. The covariance of the independent variables is equal to 0, and the covariance of a random variable, which is a common part for both components, with itself is equal to the variance: $\frac{1}{s^2}(s - m + k)\sigma^2$. If $m - k \geqslant s$, the variables S_k and S_m are independent (their intersection is empty), then their covariance is 0.

We can also estimate the conditional probability of occurrence of a L-limited segment using the normal distribution in the case $s \to \infty$:

$$\lim_{s\to\infty}\left(P\{S_i \leqslant \log_2 L \mid S_{i-1} \leqslant \log_2 L\} - \frac{\int_0^{\log_2 L}\int_0^{\log_2 L} f(x,y)dxdy}{\Phi(R)}\right) = 0, \tag{9}$$

where $f(x,y) = \frac{s}{2\pi\sigma^2\sqrt{1-\rho^2}} \times e^{-\frac{1}{2(1-\rho^2)}\left(\frac{s(x-\mu)^2}{\sigma^2} - \rho\frac{2s(x-\mu)(y-\mu)}{\sigma^2} + \frac{s(y-\mu)^2}{\sigma^2}\right)}$, $\rho = \frac{s-1}{s}$ is the correlation coefficient of the ith and $(i-1)$th segments, $R = \sqrt{s}\cdot\frac{\log_2 L - \mu}{\sigma}$.

In this case, the expected average value of the next segment, provided that the previous one is L-limited [17], is as follows:

$$\lim_{s\to\infty}\left(\mathrm{E}(S_i \mid S_{i-1} \leqslant \log_2 L) - \left[\mu - \frac{s-1}{\sqrt{s}s}\sigma\frac{\phi(R)}{\Phi(R)}\right]\right) = 0, \tag{10}$$

where ϕ is the density of the standard normal distribution, Φ is the standard normal distribution function, and $R = \sqrt{s}\cdot\frac{\log_2 L - \mu}{\sigma}$.

Let us define a random vector $\vec{S} = (S_1, S_2, \ldots, S_{N-s+1})^T$ characterizing the message with a length of N characters. The probability that all segments of a message are simultaneously L-limited tends to a multidimensional normal distribution as $s \to \infty$:

$$\begin{aligned} &P\{S_1 \leqslant \log_2 L, S_2 \leqslant \log_2 L, \ldots, S_{N-s+1} \leqslant \log_2 L\} \\ &= P\{\vec{S} \leqslant \log_2 L\} \xrightarrow{s\to\infty} \mathcal{N}(\vec{\theta}, \Sigma), \end{aligned} \tag{11}$$

where

$$\vec{\theta} = \begin{pmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{pmatrix}$$

is the vector of average values,

$$\Sigma = \begin{pmatrix} \sigma^2 & & \sigma_{S_1, S_{N-s+1}} \\ \vdots & \ddots & \vdots \\ \sigma_{S_{N-s+1}, S_1} & & \sigma^2 \end{pmatrix}$$

is the covariance matrix, at the intersection kth row and mth column of which we have the values of the covariance of the segments:

$$\begin{aligned} \sigma_{S_k, S_m} &= \mathrm{Cov}(S_k, S_m) \\ &= \begin{cases} \frac{(s-m+k)\sigma^2}{s^2} & \text{if} \quad m-k < s \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{12}$$

$\vec{S}$ is the multidimensional normal vector of dimension $N - s + 1$ whose components are the quantities S_i having a normal distribution as $s \to \infty$.

The probability of simultaneous L-limitedness of all segments of the message in the limit is subject to a multidimensional normal distribution with the corresponding covariance matrix and vector of average values.

In this case, the vector components S_k and S_m such that $m - k < s$ have nonzero covariance.

For a number of probability distributions (for example, uniform), the sum of independent variables quickly converges to the normal distribution. Traditional estimates of the error generated by the central limit theorem in the final case, such as the Berry–Essen inequality, show only rough estimates and inflated upper bounds at the mean value s (tens). This is due to the fact that, firstly, different classes of distributions converge to 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 at large s. It is known from numerical calculations that the uniform distribution quickly converges to normal, much faster than general theoretical estimates [18]. Therefore, when summing uniformly distributed random variables, it is possible already with 6–10 terms to achieve sufficient closeness to the normal law [23].

Using the normal distribution, we can approximately estimate the probability of occurrence of L-limited segments in the case of a finite value of the quantity s. Further, in the calculation examples for uniform distribution, statements 1–4 are applied, starting from the value $s \geq 10$.

5.2. Key Reuse Case

Let the encryption key be reused for M various messages. At the same time, for the first $M - 1$ messages, pairs of plaintext and ciphertext are known. For the last message, only the ciphertext is known. Then, for the unknown plaintext of the last message, it is possible to determine some information about the variants of its characters.

Cipher texts M of the messages of the same length are matched against each other. Where the encrypted characters coincide, these messages also have the same original characters, since the key sequence did not change during encryption. Thus, in all places where the encrypted characters of the last message coincide with the characters of any of the previous messages, the open characters of the unknown message are uniquely restored. For other characters, $m - M + 1$ possible values remain (m is the cardinality of the plaintext alphabet).

If the key is reused, a random indicator model can be used to describe the distribution of the number of possible values. Either a match was found with a probability p on the place of the ith character of the message, or no matches were found with a probability $1 - p$. According to experimental studies when the key was used twice, $p = 0.06$.

Let the same key sequence be used several times for different messages of the same length. In this case, the plaintexts are known for all messages except the last one. The ciphertexts of all messages are divided into s-grams (with gearing) and are compared character by character.

Let us introduce a random variable:

$$\xi_{i_j} = \begin{cases} 1 & \text{if for } i_j\text{th character} \\ & \text{at least one match is found} \\ 0, & \text{if there are no matches.} \end{cases}$$

In places of matches, the unknown character of the last message is determined from information about the open characters of previous messages. Then the number l_{i_j} of character options for all segments of an unknown message is

$$l_{i_j} = \begin{cases} 1, & \xi_{i_j} = 1 \\ m - M + 1, & \xi_{i_j} = 0, \end{cases} \tag{13}$$

where m is the cardinality of the alphabet and M is the number of messages with the same key.

Suppose that the probability $P\{\xi_{i_j} = 1\} = p$ and $P\{\xi_{i_j} = 0\} = 1 - p$, respectively, for any i and j.

The average characteristic of the ith segment of the message in case of re-encryption is called a random variable:

$$\hat{S}_i = \sum_{j=1}^{s} \xi_{i_j}. \tag{14}$$

It is clear that $\hat{S}_i$ is the number of characters per ith segment for which it was possible to restore the original sign.

The mathematical expectation and variance of a quantity $\hat{S}_i$:

$$E\hat{S}_i = sp, \quad D\hat{S}_i = sp(1 - p).$$

Let the value of the critical threshold L be fixed. Then the following statements about the distribution of the probability of occurrence of an L-limited segment in an unknown message are true.

The probability that the geometric mean of any ith segment of the unknown message does not exceed the specified boundary L, has a binomial distribution with parameters s and p:

$$\begin{aligned} &P\{\sqrt[s]{l_{i_1} \cdot \ldots \cdot l_{i_s}} \leqslant L\} \\ &= P\left\{\hat{S}_i \geqslant s - s\frac{\log_2 L}{\log_2(m - M + 1)}\right\} \\ &= 1 - \sum_{k=0}^{v} C_s^k \cdot p^k \cdot (1 - p)^{s-k} \end{aligned} \tag{15}$$

for any i, where C_s^k is the binomial coefficient, where

$$v = \left[s - s\frac{\log_2 L}{\log_2(m - M + 1)} \right].$$

The covariance between the kth and mth arbitrary segments of the message with s characters is equal to

$$\mathrm{Cov}(S_k, S_m) = \begin{cases} (s - m + k) \cdot p \cdot (1 - p) & \text{if} \quad m - k < s \\ 0, \quad \text{otherwise.} \end{cases} \tag{16}$$

For a message with N characters, consider a binomial vector:

$$\vec{S} = (S_1, S_2, \ldots, S_{N-s+1})^{\mathrm{T}},$$

where $M(s)$ is the average characteristic of the ith s-gram of the message having a binomial distribution with parameters s and p.

The probability that all message segments of length s are simultaneously L-bounded, has a multinomial distribution with a vector of mean values θ and covariance matrix Σ:

$$\vec{\theta} = (sp, sp, \ldots, sp)^{\mathrm{T}},$$

$$\Sigma = \begin{pmatrix} sp(1-p) & \ldots & \mathrm{Cov}(S_1, S_{N-s+1}) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(S_{N-s+1}, S_1) & \ldots & sp(1-p) \end{pmatrix}. \tag{17}$$

5.3. Mathematical Model of Distribution of the Number of Meaningful Texts

Let $N_s = m^s$ be a total number of all s-grams in the alphabet of cardinality m, among which there are exactly $M(s)$ meaningful s-grams. The value $M(s)$ is estimated as $2^{H_s \cdot s}$, where H_s is the entropy of s-grams per character [12, 13].

Next, we define n_i as the number of possible recovery variants for the ith segment, among which there is one true variant. Since the composition of the output sets for each unknown message character is formed by a random, equally probable selection of characters from the original alphabet, we consider the set $n_i - 1$ (of false recovery options) a sample on the set $N_s - 1$, in which there are $M(s) - 1$ meaningful texts.

Then the probability that, in a sample of $n_i - 1$ of different recovery variants, exactly r_i variants turn out to be meaningful, is described by the hypergeometric distribution [17], that is,

$$P(n_i - 1, r_i) = \frac{C_{M(s)-1}^{r_i} C_{N_s - M(s)}^{n_i - 1 - r_i}}{C_{N_s - 1}^{n_i - 1}}. \tag{18}$$

If the ith segment of the message (the cardinality of the alphabet is m characters) with s characters and average number of values L_i is recovered, then the probability of occurrence of r_i meaningful recovery variants for a given segment is

$$P(L_i^s - 1, r_i) = \frac{(2^{H_s s} - 1)!(m^s - 2^{H_s s})!(L_i^s - 1)!(m^s - L_i^s)!}{r_i!(2^{H_s s} - 1 - r_i)!(L_i^s - 1 - r_i)!(m^s - 2^{H_s s} - L_i^s + 1 + r_i)!(m^s - 1)!}. \tag{19}$$

This statement follows from formula (18). In this case, the probability of finding exactly one meaningful text (true) is estimated at the value of the parameter $r_i = 0$.

Thus, the probability of successfully recovering the ith s-gram with an average number of values L_i taking into account admissible ambiguity is defined as

$$P_{\mathrm{uniq}}(s, L_i, \beta) = \sum_{r_i = 0}^{r_{\max} - 1} P(L_i^s - 1, r_i), \tag{20}$$

where $r_{\max} = \lfloor 2^{\beta s} \rfloor$ is the maximum admissible degree of reconstruction ambiguity of an s-gram and β is the algorithm parameter.

The limit distribution of the number of meaningful texts occurs as $s \to \infty$. Then the parameters $N_s = m^s \to \infty$, $M(s) = 2^{Hs} \to \infty$, H is the limiting value of entropy H_s as $s \to \infty$. At the same time, we believe $L_i \leqslant \frac{m}{2}$. The type of marginal distribution depends on the number of meaningful texts in the sample.

Because statements 21 and 22 below describe cases of limit distribution as $s \to \infty$, they ignore the entropy of short s-gram and use the corresponding limiting value of entropy (i.e., the entropy of language).

The probability of finding exactly 1 meaningful text (true) as $s \to \infty$ is as follows:

$$\lim_{s \to \infty} \left(P(L_i^s - 1, 0) - \exp\left\{ -2\left(\frac{L_i \cdot 2^H}{m} \right)^s \right\} \right) = 0. \tag{21}$$

Using Stirling's formula and the second remarkable limit, we obtain

$$P(n_i - 1, 0) = \frac{C_{N_s - M(s)}^{n_i - 1}}{C_{N_s - 1}^{n_i - 1}} = \frac{(N_s - M(s))!(N_s - n_i)!}{(N_s - M(s) - n_i + 1)!(N_s - 1)!}$$

$$\to \frac{(N_s - M(s))^{N_s - M(s)} \cdot (N_s - n_i)^{N_s - n_i}}{(N_s - M(s) - n_i + 1)^{N_s - M(s) - n_i + 1} \cdot (N_s - 1)^{N_s - 1}} \to \exp\left\{-2\left(\frac{L_i \cdot 2^H}{m}\right)^s\right\}.$$

The probability of getting $r_i = \lfloor 2^{\beta s} \rfloor$ meaningful texts, among which one is true, when restoring ith segment of length $s \to \infty$ has the form:

$$\lim_{s \to \infty}\left(P(L_i^s - 1, r_i = \lfloor 2^{\beta s} \rfloor - 1) - \frac{1}{\sqrt{\pi}} 2^w\right) = 0, \qquad (22)$$

where

$$w = \left(\left(H - \beta + \log_2 \frac{L_i}{m}\right) s + \log_2 e\right) \cdot 2^{\beta s} - \frac{\beta}{2} s - \frac{1}{2}.$$

When using statements above 5 for numerical calculations, the parameter value s must satisfy the condition $H_s \approx H$. For the Russian language $s \geqslant 50$, $H \approx 0.78$.

If the admissible number of meaningful texts depends on s, then $r_i = \lfloor 2^{\beta s} \rfloor \to \infty$. Assuming the parameters $\beta < 1$; $L_i \leqslant \frac{m}{2}$:

$$P(L_i^s - 1, r_i = \lfloor 2^{\beta s} \rfloor - 1) = \frac{C_{M(s)-1}^{r_i} C_{N_s - M(s)}^{n_i - 1 - r_i}}{C_{N_s - 1}^{n_i - 1}}$$

$$\to \frac{\sqrt{N_s - M(s)}\sqrt{N_s - n_i}}{\sqrt{2\pi}\sqrt{N_s - M(s) - n_i}\sqrt{N_s}} \times \frac{(N_s - M(s))^{N_s - M(s)}(N_s - n_i)^{N_s - n_i}}{(N_s - M(s) - n_i)^{N_s - M(s) - n_i} N_s^{n_i}}$$

$$\to \frac{1}{\sqrt{2\pi}\sqrt{r_i}}\left(\frac{M(s) \cdot n_i}{r_i \cdot N_s}\right)^{r_i} e^{r_i} = \frac{1}{\sqrt{\pi}} 2^w,$$

where $w = \left(\left(H - \beta + \log_2 \frac{L_i}{m}\right) s + \log_2 e\right) \cdot 2^{\beta s} - \frac{\beta}{2} s - \frac{1}{2}.$

Section 8 provides some numerical calculations of the proved statements.

6. EFFICIENCY OF THE ALGORITHM

In practice, the efficiency of the algorithm can be considered as the total share of information that was recovered at the output of the communication channel. In theoretical estimates, this share is determined by the probability of recovery of a L-limited segment of a given length, that is, the joint probability of the appearance of such a segment in a message and the success of its recovery, taking into account acceptable ambiguity.

Then we estimate the total probability of recovering message segments, the average value of the variants of which does not exceed the given critical threshold L, ignoring the degree of dictionary coverage, as

$$P_{\text{recov}}(s, L, \beta) = P_{\text{occur}}(s, L) \cdot \sum_{L_i < L} P_{\text{uniq}}(s, L_i, \beta), \qquad (23)$$

where $P_{\text{occur}}(s, L)$ is the probability of occurrence of the ith segment of length s with an average number of options not exceeding L; $P_{\text{uniq}}(s, L_i, \beta)$ is the probability that recovery of the ith segment with the average number of variants L_i does not exceed the admissible ambiguity.

Taking into account incomplete coverage of the dictionary used, the total share of recovered information at the output of the communication channel has the form:

$$\pi = \tau_s \cdot P_{\text{recov}}(s, L, \beta), \qquad (24)$$

where τ_s is the degree of dictionary coverage of s-grams.

If the total share of recovered information at the output of the communication channel exceeds the set critical value $\pi > \pi_0$, then this communication channel is estimated as unprotected.

7. COMPUTATIONAL COST OF THE ALGORITHM

The costly stages of compiling and sorting dictionaries are preliminary and are not included in the calculation of the computational complexity of the algorithm itself. Thus, the complexity of the algorithm is determined by the implementation complexity of the stage of recovering an L-limited segment.

Let L be a critical threshold for selecting message segments and suppose that s is the length of the message segment to be restored, D_s is the volume of the dictionary of s-grams. Then the number of recovery variants of the L-limited segment checked for presence in the dictionary does not exceed L^s.

The asymptotic recovery complexity of an L-bounded segment when implementing a the binary algorithm of dictionary search is given by

$$Q(s, L) = O(L^s \cdot \log_2 D_s). \qquad (25)$$

The total number of L-limited segments in a message with N characters is given in statement 6.

8. NUMERICAL ESTIMATES OF ALGORITHM PARAMETERS FOR THE RUSSIAN LANGUAGE

Let us consider an algorithm for restoring a message in Russian with a fixed plaintext alphabet, i.e., 35 characters (32 letters of the Cyrillic alphabet, space, period, and comma).

Table 2. Entropy of the Russian language

Length of the segment, s	10	15	20	25	More than 50
Entropy, H_s	2.49	1.80	1.41	1.16	0.78

Table 3. Algorithm parameters

Cardinality of the alphabet of code characters, m	35
Parameter β	0.1
The degree of admissible ambiguity, $r_{\max}$	See Table 1
Length of the restored segment, s	10–25
The critical boundary for selecting segments, L	8–24
Entropy of s-gram, H_s	See Table 2
Probabilities of occurrence of output character values, p_k	$p_k = \frac{1}{35}, \forall k = 1,\dots,35$

Table 4. Most probable number of meaningful texts

$L\|s$	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 5. Theoretical share of reconstructed segments

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	0.935
<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

Entropy estimates were used in the calculations of s-grams of the Russian language [12]:

Because the entropy value stabilizes for segments longer than 50 characters, the value of 0.78 bits per character is taken as the maximum entropy level.

The studied parameters of the algorithm are given in Table 3.

Theoretically, the most probable number of meaningful texts (taking into account the true one) in Russian that are found when reconstructing a message segment is given in Table 4.

Consequently, for the Russian language, restoring 10-grams with an average number of values greater than 8–9 characters is practically meaningless. For 15-grams, the maximum recovery efficiency is for the value L of no more than 12–13 characters.

As the length of the message segment increases, the degree of ambiguity in plaintext reconstruction decreases. However, for long sections of text, the process of compiling search dictionaries becomes very difficult. The simulation allows choosing the optimal algorithm parameter—the length of the message segment to be restored—based on the most effective ratio between the degree of ambiguity of plaintext recovery and the degree of coverage of the corresponding dictionary.

Obviously, this parameter is the minimum length of an s-gram, in which the probability of restoring a segment approaches one for any average number of character variants. Based on the probabilities of the hypergeometric distribution, the length of such a minimum s-gram is determined for the Russian language (with an alphabet of 35 characters) to be 16 characters.

For a number of cases of a specific type of probability distribution $p_k = P(l_{i_j} = k)$ (number of values of

Table 6. Dictionaries of s-grams

Length of the segment, s	10	15	20	25
Dictionary size, D_s	7.9×10^5	9.5×10^5	9.8×10^5	9.9×10^5
Entropy of s-gram, H_s	1.96	1.32	0.99	0.80

Table 7. Evaluating the algorithm efficiency

L	Experimental evaluation				Theoretical assessment			
	10	15	20	25	10	15	20	25
8	0.019	0.012	0.004	0.001	0.019	0.006	0.002	0.001
11	0.049	0.163	0.139	0.116	0.069	0.168	0.136	0.110
16	0.052	0.483	0.761	0.808	0.069	0.525	0.762	0.782

the i_jth input character), it is easy to obtain numerical estimates of the algorithm's efficiency.

Let, for example, the number of values for all segments be distributed independently and by chance equally probable from 1 to m, that is, it takes values with the same probabilities for any $k = 1, 2, \ldots, m$. The critical threshold value L is fixed.

Theoretical assessment of the efficiency of the algorithm for the Russian language (the share of reconstructed segments in a message ignoring the degree of dictionary coverage) for different parameter values s and L when the number l_{i_j} of values for all segments is distributed evenly from 1 to 35 are shown in Table 5.

Thus, as the average number of meanings of unknown characters for a segment in Russian with a length of at least 16 characters increases, the probability of its recovery grows, if we take into account the permissible degree of ambiguity.

9. EXPERIMENTAL STUDY OF THE ALGORITHM

As part of the confirmation of the constructed model, a comparison of theoretical estimates of the effectiveness of the algorithm with experimental results is carried out. To carry out the experiment, we use a software implementation[2] of the proposed algorithm and the created corpus of journalistic texts in the Russian language with a volume of 1 million characters,[3] described in detail in [12, 13].

Based on a text corpus we programmatically[4] create the dictionaries of s-grams used when restoring a message (Table 6).

We sequentially consider the recovery of message segments of 10, 15, 20, and 25 characters in length. The total message length is $N = 1000$ characters. For different values of the critical threshold L, the relative proportion of successfully restored segments is determined.

In addition, to make a comparison, the theoretical efficiency of the algorithm is evaluated under the same parameters. When calculating, we assumed the number of meaningful texts with s characters to be equal to the size of the corresponding dictionary of s-grams (possible incomplete coverage of dictionaries is ignored).

10. CONCLUSIONS

We considered the problem of restoring a message transmitted in a discrete communication channel, when the output contains information about the possible values of the source text, which arose as a result of violations in the operation of elements of the information security system. Unlike previous approaches, we proposed an algorithm for recovering individual s-grams of the message when the complete recovery based on the available information is impossible. The proposed procedure consists of several stages, including preliminary compilation of dictionaries of s-grams. We assessed the effectiveness of the proposed algorithm within the framework of an equiprobable model, with the help of which we determined the optimal parameters of the algorithm for the Russian language and the criterion for the insecurity of the communication channel. Theoretical estimates of efficiency were confirmed by experimental testing of the constructed algorithm model.

During the study, we examined various types of distributions of the number of possible characters in the output message. As a result, we found that, as the length of a text segment increases, the probability of finding L-limited s-grams potentially suitable for recovery approaches a normal distribution regardless of the initial number of values. In the case of key reuse, this probability is described by a binomial distribution. In addition, we found that the normal distribution allows well approximating the initial probabilities even for small values of length of the s-gram.

When studying the degree of ambiguity in the recovery of text segments resulting from the application of the proposed algorithm, we used the hypergeometric distribution. Using it, we obtained numerical calculations of the probability that an admissible number of meaningful texts occur when recovering one segment. In addition, we found the asymptotics of the hypergeometric distribution for large segment lengths.

The overall efficiency of the algorithm was considered as the average proportion of information recovered in an unknown message. To assess the complexity of implementing the algorithm, we determined its computational cost, which mainly depends on the complexity of searching through the dictionary. As a result of studying the recovery efficiency for the Russian language (ignoring the degree of dictionary coverage), we determined the optimal parameters of the algorithm (Table 8).

[2] The result of intellectual activity was created at the HSE University. https://github.com/Nastasian/recovery.

[3] https://github.com/Nastasian/entropy/releases/download/v2/Russian_political_news_corpus.txt.

[4] The result of intellectual activity was created at the HSE University. https://github.com/Nastasian/entropy.

Table 8. Optimal algorithm parameters for the Russian language

Length of the message segment to be restored is $s = 16$ characters	Recovery efficiency	
Maximum number of possible values for a character of an s-gram	14 characters	50%
	16 characters	75%
	18 characters	90%

The main application of the results is in cryptography when analyzing the strength of cryptographic algorithms, in particular, in situations of incorrect selection of a key or its reuse in symmetric encryption algorithms [21], as well as in cases of various leaks of information about the characters of the original message.

FUNDING

This work was supported by ongoing institutional funding. No additional grants to carry out or direct this particular research were obtained.

CONFLICT OF INTEREST

The author of this work declares that she has no conflicts of interest.

REFERENCES

1. I. Gorbenko, A. Kuznetsov, M. Lutsenko, and D. Ivanenko, “The research of modern stream ciphers,” in *Proceedings of the 4th International Scientific-Practical Conference Problems of Infocommunications, Science and Technology* (*PIC S&T*) (IEEE, 2017), pp. 207–210.
2. J. P. Aumasson, *Serious Cryptography: A Practical Introduction to Modern Encryption* (No Starch, San Francisco, 2017).
3. S. Rubinstein-Salzedo, *Cryptography* (Springer, Cham, 2018).
4. M. Barakat, C. Eder, and T. Hanke, *An Introduction to Cryptography* (Technische Univ. Kaiserlautern, 2018).
5. G. C. Kessler, “An overview of cryptography” (2020). www.garykessler.net/library/crypto.html
6. B. V. Gnedenko, E. K. Belyaev, and A. D. Solov’ev, *Mathematical Methods in Reliability Theory* (Librokom, Moscow, 2019) [in Russian].
7. A. Savinov and V. Ivanov, “Analysis of solutions to problems related type I and II errors arising in keyboard dynamics recognition systems,” Vestn. Volzhsk. Univ. im. V.N. Tatishcheva, No. 18 (2011).
8. A. V. Babash, “Attacks on the random gamming code,” Math. Math. Model., No. 6, 35–38 (2020).
9. A. V. Babash, E. K. Baranova, A. A. Lyutina, A. A. Murzakova, E. A. Murzakova, D. M. Ryabova, and E. S. Semis-Ool, “About text noise bounds with text content saving: Applications to cryptography,” Vopr. Kiberbezopasnosti, No. 1 (35), 74–86 (2020).
10. B. Coecke, T. Fritz, and R. W. Spekkens, “A mathematical theory of resources,” Inf. Comput. **250**, 59–86 (2016).
11. A. G. Malashina and A. B. Los’, “Construction and analysis of Russian language models for cryptographic algorithm research,” Chebyshev. Sb. **23** (2), 151–160 (2022).
12. A. G. Malashina, “Development of instrumental tools for studying information characteristics of a natural language,” Promyshlennye ASU i Kontrollery, No. 2, 9–15 (2021).
13. A. Malashina, “The combinatorial analysis of n-gram dictionaries, coverage and information entropy based on the web corpus of English,” Balt. J. Mod. Comput. **9** (3), 363–376 (2021).
14. U. Nurmukhamedov, “Lexical coverage and profiling,” Lang. Teach. **52** (2), 188–200 (2019).
15. T. S. Juzek, “Using the entropy of N-grams to evaluate the authenticity of substitution ciphers and Z340 in particular,” *Proceedings of the 2nd International Conference on Historical Cryptology, HistoCrypt, 2019, Linköping Electronic Conference Proceedings* (2019), Vol. 158, pp. 177–125.
16. M. Morzy, T. Kajdanowicz, and P. Kazienko, “On measuring the complexity of networks: Kolmogorov complexity versus entropy,” Complexity **2017**, 3250301 (2017). https://doi.org/10.1155/2017/3250301
17. W. H. Greene, “Limited dependent variables—truncation, censoring and sample selection,” in *Econometric Analysis* (Prentice Hall, 2008), pp. 833–902.
18. V. M. Zolotarev, *Modern Theory of Summation of Random Variables* (De Gruyter, Berlin, 2011).
19. M. A. Taha, N. M. Sahib, and T. M. Hasan, “Retina random number generator for stream cipher cryptography,” Int. J. Comput. Sci. Mobile Comput. **8** (9), 172–181 (2019).
20. J. Poonam and S. Brahmjit, “RC4 encryption—a literature survey,” Procedia Comput. Sci. **46**, 697–705 (2015).
21. V. K. Pachghare, *Cryptography and Information Security* (PHI Learning, Delhi, 2019).
22. C. E. Shannon, “A mathematical theory of communication,” Bell Syst. Tech. J. **27** (3), 379–423 (1948).
23. S. A. Aivazyan, *Applied Statistics: Basics of Modeling and Primary Processing: Reference Book* (Finansy i Statistika, Moscow, 1983) [in Russian].

Publisher’s Note. Pleiades Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.