ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2023, том 514, № 2, с. 138–149

УДК 003.26.09

О ВОЗМОЖНОСТИ ВОССТАНОВЛЕНИЯ ОТРЕЗКОВ СООБЩЕНИЯ ПО ИНФОРМАЦИИ О ЗНАЧЕНИЯХ ИСХОДНЫХ СИМВОЛОВ

© 2023 г. А. Г. Малашина[1,*]

Представлено академиком РАН А.Л. Семеновым
Поступило 04.09.2023 г.
После доработки 15.09.2023 г.
Принято к публикации 24.10.2023 г.

В целях обеспечения защищенного информационного обмена в каналах связи необходимо предварительное исследование корректности работы соответствующих систем защиты информации. Несмотря на то что используемые в таких системах математические алгоритмы корректны и теоретически обеспечивают правильные статистические свойства выходного потока по сравнению с входным, на этапе реализации (программирования) данных алгоритмов защиты или на этапах сборки конечного оборудования (использования аппаратных средств, внесения настроек) и его эксплуатации в реальных условиях возможно внесение искажений, которые нарушают работу тех или иных элементов средств защиты информации (например, датчика случайных чисел). В результате по характеру передаваемого сигнала становится возможным определить, что выходной поток по ряду характеристик устойчиво отличается от идеального шифрованного потока, который в теории должен был исходить от оборудования и появляться на выходе канала связи. В данной ситуации необходимо понимание того, насколько внесение тех или иных искажений влияет на степень защищенности создаваемой системы. С этой целью приводится описание параметров различных источников сообщений, которые моделируют получение выходного потока с искажениями. При этом степень защищенности соответствующего канала связи предлагается определить с помощью оценки доли входного потока, которую удается восстановить из выходного, используя побочную информацию, возникающую в результате внесения соответствующих искажений в работу системы.

Ключевые слова: осмысленные тексты, атака по словарям, каналы связи, каналы перехвата
DOI: 10.31857/S2686954323601963, **EDN:** CVVQXN

1. ВВЕДЕНИЕ

1.1. Введение в проблему

В процессе реализации и эксплуатации некоторых систем защиты информации возможно внесение нарушений в работу различных элементов оборудования (например, датчика случайных чисел), искажающих корректный результат применения математических алгоритмов шифрования.

Подобные нарушения могут привести к случаям [1–3, 20, 21]:

• уменьшения мощности алфавита ключа или неравновероятного распределения символов ключа (при применении шифра гаммирования);

• утечкам информации о символах исходного сообщения;

• повторного использования ключа шифрования.

Таким образом, в результате искаженной работы математического алгоритма шифрования на выходе канала связи появляется побочная информация о символах входного сообщения. Обладая данной информацией и предполагая осмысленность[1] исходного сообщения в рассматриваемом языке, можно построить алгоритм для восстановления исходного текста или его отдельных частей.

Полное восстановление возможно, если количество предполагаемых значений исходных символов сообщения ограничено, при этом вероятность появления среди них истинного знака близка к единице. Тогда для каждого входного символа фиксируется множество всех его возможных значений. Восстановление исходного текста представляет собой поиск осмысленного текста среди всех возможных комбинаций [4, 5].

[1] *Национальный исследовательский университет "Высшая школа экономики" (МИЭМ, ИСИЭЗ), Москва, Россия*
**E-mail: amalashina@hse.ru*

[1] Под осмысленностью текста подразумевается допустимость его существования в языке.

Такой подход может привести к потере истинного варианта восстановления, вероятность которой оценивается исходя из введенной теоретико-вероятностной модели исследуемого процесса [6–8]. С увеличением количества значений на выходе канала связи восстановление исходного текста становится затруднительным из-за высокой степени неопределенности выбора осмысленного текста, возникающей в процессе поиска подходящих вариантов. То есть трудность состоит в выборе истинного исходного текста среди большого количества восстановленных осмысленных текстов.

Степень неоднозначности восстановления r – это математическое ожидание числа возможных осмысленных текстов, которые могут быть получены путем применения алгоритма восстановления [2]. Значение r зависит от длины текста s и позволяет оценить эффективность восстановления исходного текста. Цель процедуры восстановления – получить подлинный осмысленный текст.

Таким образом, когда количество символов на выходе канала невелико, осмысленный текст может быть построен единственным способом, так как все остальные комбинации окажутся текстом случайной структуры. Однако по мере увеличения числа значений такой подход приводит к нахождению более одного возможного варианта восстановления. В этом случае невозможно определить, какой из найденных текстов является исходным сообщением, не обладая дополнительной информацией. Поэтому применение данной процедуры для восстановления сообщения *целиком* становится неэффективным.

Тем не менее возникает вопрос о возможности восстановления *отдельных частей* неизвестного сообщения. Если соответствующие значения наборов знаков являются реализацией случайной величины, то при достаточной длине сообщения могут появляться редкие благоприятные события, когда на отдельных участках оказываются наборы небольшой длины. На таких отрезках можно построить алгоритм определения искомой части неизвестного текста, например путем предварительного составления словарных величин соответствующей длины.

1.2. Постановка задачи

Пусть при передаче знаков входного сообщения $x_1,\ldots,x_N$, выбираемых из алфавита A мощностью m, на выходе канала связи наблюдаются множества символов различной длины $\{x_i\}_{l_i} = (x_i^{(1)},\ldots,x_i^{(l_i)})$, $x_i^{(j)} \in A \ \ \forall i = 1,\ldots,N,\ j = 1,\ldots,l_i$. Значения величин l_i предполагаются таковыми, что полное восстановление сообщения невозможно. При этом задача состоит в поиске участков сообщения длиной s символов с относительно небольшими значениями l_i и попытках восстановить на них отдельные части передаваемого сообщения. В рамках определенной теоретико-вероятностной модели появления наборов $\{x_i\}_{l_i}$ и моделей словарей требуется оценить среднюю долю восстановленного текста исходного сообщения.

2. МАТЕМАТИЧЕСКАЯ МОДЕЛЬ КАНАЛА СВЯЗИ

Приведем описание канала связи, моделирующего получение выходного потока при внесении нарушений в работу элементов систем защиты информации.

Предположим, для каждого символа сообщения оказалось возможным построить определенный набор значений, среди которых присутствует истинный искомый символ. Количество таких возможных вариантов может иметь различные вероятностные распределения.

Рассмотрим соответствующий дискретный канал связи без памяти [9, 10]. Пусть в качестве входного и выходного конечных алфавитов выступает множество символов A ($|A| = m < \infty$). Каждому i-му входному символу сообщения x_i соответствует множество $\{x_i\}_{l_i}$ значений мощностью l_i.

Зададим вероятности $p_k^{(i)} = P(l_i = k)$ появления на выходе множества $\{x_i\}_k$, состоящего из k значений; $k = 1,\ldots,m$; $\sum_{k=1}^{m} p_k^{(i)} = 1$.

Далее определим, что состав множества значений $\{x_i\}_k$ для каждого входного символа образуется случайным и равновероятным выбором из всех упорядоченных последовательностей длины k, построенных без повторения знаков в исходном алфавите.

Таким образом, модель канала определяется:

- алфавитом A кодовых символов на входе и выходе мощностью m;
- множеством входных сообщений X длиной N символов, где $x = (x_1, x_2,\ldots, x_N) \in X$ является входным сообщением, $x_i \in A$, $\forall i = 1,\ldots,N$;
- множеством значений выходных символов $(\{x_1\}_{l_1},\{x_2\}_{l_2},\ldots,\{x_N\}_{l_N})$, где l_i – мощность множества $\{x_i\}_{l_i}$, $\forall i = 1,\ldots,N$;
- вероятностями $p_k^{(i)}$ появления выходных множеств $\{x_i\}_k$ для $k = 1,\ldots,m$, при этом состав множества $\{x_i\}_k$ образуется путем выбора каждого знака из алфавита A по урновой схеме без возвращения, $\forall i = 1,\ldots,N$.

Далее предлагается алгоритм восстановления отдельных частей неизвестного сообщения, передаваемого в данном канале связи, и рассматривается случай фиксированного произвольного вероятностного распределения $p_k^{(i)} = p_k$ в канале связи для всех $i = 1,\ldots,N$.

3. ОПИСАНИЕ АЛГОРИТМА

3.1. Об s-граммах сообщения

Для реализации алгоритма восстановления сообщение разделяется на отдельные s-граммы. Под s-граммой понимается последовательность из s символов текста, которые выбираются "с зацеплением", т.е. для каждой следующей s-граммы осуществляется сдвиг вправо на один символ.

Номер s-граммы в сообщении определяется по порядковому номеру ее первого символа. То есть i-я s-грамма обозначается как $(x_i, x_{i+1},\ldots, x_{i+s-1})$. При рассмотрении отдельной s-граммы вводится внутренняя нумерация ее символов. Например, для i-й s-граммы: $(x_{i_1}, x_{i_2},\ldots, x_{i_s})$. Соответствующее число значений для j-го символа i-й s-граммы обозначается как l_{i_j}. Общее количество s-грамм в сообщении длиной N символов равно $N - s + 1$.

3.2. Алгоритм восстановления s-грамм

Вход алгоритма: общая длина неизвестного сообщения (N), длина восстанавливаемых участков (s), информация о возможных значениях каждого i-го знака сообщения ($\{x_i\}_{l_i}$) и их количестве (l_i), критическая граница отбора участков (L), параметр (β), предварительно созданный словарь s-грамм и его мощность (D_s).

Алгоритм включает следующие этапы:

1. *Выбор подходящих отрезков сообщения*. Для восстановления необходимо выбрать те s-граммы сообщения, для которых известно небольшое среднее количество значений.

Критерий выбора *s*-грамм: среднее геометрическое значение L_i числа возможных символов для соответствующей i-й s-граммы не должно превышать заданной критической границы L для любого i:

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

где s – длина сегмента текста (s-граммы), l_{i_j} – число вариантов для j-го символа в i-й s-грамме, L_i – среднее число вариантов для i-й s-граммы, L – критическая граница отбора сегментов.

2. *Построение вариантов восстановления s-граммы*. Поиск всех возможных комбинаций, которые могут быть построены с использованием известной информации о значениях знаков s-граммы.

3. *Выбор осмысленных вариантов восстановления*. Если составленный текст s-граммы присутствует в словаре, то он считается осмысленным и принимается как возможный вариант восстановления s-граммы сообщения. Если совпадений со словарем не обнаружено, данный вариант восстановления отклоняется.

4. *Восстановление s-граммы сообщения*. В соответствии с числом осмысленных s-грамм, которое удалось построить, выбранный отрезок сообщения считается восстановленным или невосстановленным.

Критерий восстановления *s*-граммы: степень неопределенности восстановления (число вариантов восстановления для одного отрезка) не должна превышать $r \leqslant r_{\max} = \lfloor 2^{\beta s} \rfloor$, где s – длина сегмента сообщения, β – численный параметр меньше 1. На практике часто используется значение $\beta = 0.1$ (табл. 1).

Таблица 1. Допустимая степень неоднозначности восстановления отрезка, $\beta = 0.1$

Длина текста, s	10	15	16	20	25
Количество возможных вариантов, $r_{\max}$	2	2	3	4	5

Выход алгоритма: варианты восстановления отдельных s-грамм сообщения, число восстановленных s-грамм.

Этап создания словарей коротких s-грамм заданного покрытия не входит в данный алгоритм. Словари создаются предварительно на основе соответствующего текстового корпуса [12, 13] (подробнее в разделе 4).

При численных оценках рассматриваются следующие параметры алгоритма:

1. Длина отрезка текста s: 10–25 символов.

2. Критическая граница L: 8–16 символов.

3. Параметр $\beta = 0.1$.

Основной параметр алгоритма, который требуется определить, – средняя доля восстановленной информации на выходе при заданном вероятностном распределении в канале связи.

4. СЛОВАРИ *S*-ГРАММ

Словари представляют собой набор s-грамм, расположенных в алфавитном порядке без повторения. Процесс создания словаря состоит из извлечения всех s-грамм из некоторого языкового корпуса и удаления повторяющихся s-грамм перед сортировкой.

Поскольку словари составляются на основе языкового корпуса ограниченного объема, их покрытие является неполным. Это означает, что не все существующие s-граммы языка оказываются включены в данный словарь. То есть возможно появление ошибок, когда существующая s-грамма отсутствует в словаре и отбрасывается как недопустимая в данном языке.

Таким образом, степень покрытия τ_s словаря s-грамм – это отношение объема построенного словаря к общему количеству существующих s-грамм в языке [14]:

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

где D_s – объем словаря s-грамм, $M(s)$ – число осмысленных s-грамм в языке.

Проблема определения степени покрытия заключается в том, что точное количество всех осмысленных s-грамм в языке неизвестно, особенно для больших порядков, хотя асимптотическую оценку можно получить, используя теоретико-вероятностную модель Шеннона [22]. Для приблизительной оценки степени покрытия словарей s-грамм могут использоваться и другие методы: эмпирическая проверка [11, 14], оценка на основе количества однократно встречаемых s-грамм [13] и др.

Словари s-грамм представляют собой модель открытого текста, в которой сообщение рассматривается как последовательность s-грамм из словаря. В предлагаемом алгоритме данные словари рассматриваются как критерий распознавания открытого текста.

Неполнота покрытия может привести к потере истинного варианта восстановления s-граммы, если ее не окажется в используемом словаре. Поэтому τ_s является параметром, влияющим на общую вероятность успешного восстановления отрезка сообщения в канале связи на 4 шаге алгоритма.

На основе объема словаря может быть вычислена энтропия s-граммы в расчете на один символ [12, 13]:

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

Значения энтропии s-грамм используются для оценки количества осмысленных текстов заданной длины в языке [15, 16].

Предельное значение H_s принимается за энтропию языка:

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

5. МАТЕМАТИЧЕСКИЕ СВОЙСТВА АЛГОРИТМА

5.1. Вероятность появления L-ограниченных отрезков

Количество возможных значений l_{i_j} для неизвестного i_j-го символа является дискретной случайной величиной, принимающей целые значения от 1 до m с вероятностями $p_k = P(l_{i_j} = k)$ для любого i_j. При этом случайные величины l_{i_j} для всех отрезков распределены одинаково и независимо. Значение критической границы L фиксировано; значение m – конечное.

Отрезок сообщения (s-грамма) с номером i называется L-ограниченным, если:

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

Средней характеристикой i-го отрезка сообщения называется случайная величина:

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

Вероятность появления отрезка сообщения, отбираемого для восстановления, определяется вероятностью его L-ограниченности и может быть выражена через величину S_i:

$$\begin{aligned} P_{\text{появл}}(s, L) &= P\{\sqrt[s]{l_{i_1} \cdot l_{i_2} \cdot \ldots \cdot l_{i_s}} \leq L\} = \\ &= P\{S_i \leq \log_2 L\}. \end{aligned} \tag{5}$$

Поскольку для любого l_{i_j} верно, что $\mu = \mathrm{E}(\log_2 l_{i_j}) = \sum_{k=1}^{m} p_k \log_2 k$, $\sigma^2 = \mathrm{D}(\log_2 l_{i_j}) = \sum_{k=1}^{m} p_k \log_2^2 k - \left(\sum_{k=1}^{m} p_k \log_2 k\right)^2$, математическое ожидание и дисперсия величины S_i равны $\mathrm{E}S_i = \mu$ и $\mathrm{D}S_i = \frac{\sigma^2}{s}$ для любого i.

Пусть длина отрезка сообщения $s \to \infty$. Тогда согласно центральной предельной теореме [17]:

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

для любого i, где $R = \sqrt{s}\cdot\frac{\log_2 L - \mu}{\sigma}$, Φ – функция стандартного нормального распределения.

Предельное распределение из утверждения 1 может быть использовано для приближенной оценки вероятности появления L-ограниченных отрезков в сообщении в случае конечного значения величины s:

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

Так как сообщение имеет длину N символов, количество последовательных отрезков длины s равно $N - s + 1$. Ожидаемое количество L-ограниченных отрезков в сообщении равно:

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

Величины S_{i-1} и S_i соседних отрезков являются зависимыми (имеют непустое пересечение), так как имеют $s - 1$ общих компонент. Определим степень их корреляционной зависимости, показывающую долю совпадающих элементов в соседних отрезках.

а) Коэффициент корреляции двух соседних отрезков длиной s символов равен:

$$\rho = \frac{s-1}{s}. \quad (7)$$

б) Коэффициент корреляции k-го и m-го ($k < m$) произвольных отрезков длиной s символов равен:

$$\rho = \begin{cases} \dfrac{s-m+k}{s}, & \text{если} \quad m-k < s \\ 0, & \text{иначе} \end{cases} \quad (8)$$

а) Заметим, что $l_{i-1_j} = l_{i_{j-1}}$. Поскольку величины S_{i-1} и S_i имеют непустое пересечение, *коэффициент корреляции* их средних величин равен:

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

б) Заметим, что S_k является линейной комбинацией случайных величин. Обобщение этой формулы для двух произвольных отрезков производится аналогично с учетом свойства ковариации: если $m - k < s$, то S_k и S_m имеют непустое пересечение, при этом у каждой из двух компонент есть $m - k$ различных независимых слагаемых, а остальные $s - m + k$ слагаемых совпадают. Ковариации независимых величин равны 0, а ковариация случайной величины, являющейся общей частью для обеих компонент, с собой равна дисперсии: $\frac{1}{s^2}(s - m + k)\sigma^2$. Если $m - k \geqslant s$, величины S_k и S_m являются независимыми (их пересечение пусто), поэтому их ковариация равна 0.

Условная вероятность появления L-ограниченного отрезка также может быть оценена с помощью нормального распределения в случае $s \to \infty$:

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

где $f(x,y) = \dfrac{s}{2\pi\sigma^2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left(\frac{s(x-\mu)^2}{\sigma^2} - \rho\frac{2s(x-\mu)(y-\mu)}{\sigma^2} + \frac{s(y-\mu)^2}{\sigma^2}\right)}$, $\rho = \dfrac{s-1}{s}$ – коэффициент корреляции i-го и $(i-1)$-го отрезков, $R = \sqrt{s} \cdot \dfrac{\log_2 L - \mu}{\sigma}$.

При этом ожидаемая средняя величина следующего отрезка при условии L-ограниченности предыдущего [17]:

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

где ϕ – плотность стандартного нормального распределения, Φ – функция стандартного нормального распределения, $R = \sqrt{s} \cdot \dfrac{\log_2 L - \mu}{\sigma}$.

Определим случайный вектор $\vec{S} = (S_1, S_2, \ldots, S_{N-s+1})^T$, характеризующий сообщение длиной N

символов. Вероятность, что все отрезки сообщения одновременно являются L-ограниченными, стремится к многомерному нормальному распределению при $s \to \infty$:

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

где $\vec{\theta} = \begin{pmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{pmatrix}$ – вектор средних значений, $\Sigma = \begin{pmatrix} \sigma^2 & & \sigma_{S_1,S_{N-s+1}} \\ \vdots & \ddots & \vdots \\ \sigma_{S_{N-s+1},S_1} & & \sigma^2 \end{pmatrix}$ – ковариационная матрица, на пересечении k-й строчки и m-го столбца которой стоят значения ковариации отрезков:

$$\sigma_{S_k,S_m} = \mathrm{Cov}(S_k, S_m) = \\ = \begin{cases} \dfrac{(s-m+k)\sigma^2}{s^2}, & \text{если } m-k < s \\ 0, & \text{иначе} \end{cases}. \quad (12)$$

$\vec{S}$ – многомерный нормальный вектор размерности $N - s + 1$, компонентами которого являются величины S_i, имеющие нормальное распределение при $s \to \infty$.

Вероятность одновременной L-ограниченности всех отрезков сообщения в пределе подчиняется многомерному нормальному распределению с соответствующими ковариационной матрицей и вектором средних значений.

При этом ненулевую ковариацию имеют компоненты вектора S_k и S_m, такие что $m - k < s$.

Для ряда вероятностных распределений (например, равномерного) сумма независимых величин быстро сходится к нормальному распределению. Традиционные оценки погрешности, порождаемые центральной предельной теоремой в конечном случае, такие как неравенство Берри-Эссена, при среднем значении s (десятки) показывают лишь грубые оценки и завышенные верхние границы. Это связано с тем, что, во-первых, разные классы распределений сходятся к нормальному с разной скоростью, а оценка погрешности дается общая для любой структуры распределения. Во-вторых, неравенство Берри-Эссена использует оценку погрешности в виде отношения третьего момента к корню из числа слагаемых. Такая форма ошибки будет давать довольно точную оценку только на больших s. Из численных расчетов известно, что равномерное распределение быстро сходится к нормальному, намного быстрее общих теоретических оценок [18]. Поэтому при суммировании равномерно распределенных случайных величин уже при 6–10 слагаемых удается добиться достаточной близости к нормальному закону [23].

С помощью нормального распределения можно приближенно оценить вероятность появления L-ограниченных отрезков в случае конечного значения величины s. Далее в расчетных примерах для равномерного распределения утверждения 1–4 применяются, начиная со значения $s \geq 10$.

5.2. Случай повторного использования ключа

Пусть ключ шифрования был использован повторно для M различных сообщений. При этом для первых $M - 1$ сообщений известны пары открытого и шифрованного текстов. Для последнего сообщения известен только шифрованный текст. Тогда для неизвестного открытого текста последнего сообщения возможно определить некоторую информацию о вариантах его знаков.

Шифрованные тексты M сообщений одинаковой длины сопоставляются друг с другом. В местах совпадения шифрованных символов у данных сообщений совпадают и исходные символы, так как ключевая последовательность при шифровании не менялась. Таким образом, во всех местах совпадения шифрованных символов последнего сообщения с символами любого из предыдущих сообщений открытые символы неизвестного сообщения восстанавливаются однозначно. Для остальных знаков остается $m - M + 1$ возможных значений (m – мощность алфавита открытого текста).

В случае повторного использования ключа для описания распределения числа возможных значений можно воспользоваться моделью случайных индикаторов. Либо на месте i-го символа сообщения найдено совпадение с вероятностью p, либо совпадений не найдено с вероятностью $1 - p$. По данным экспериментальных исследований при двукратном использовании ключа $p = 0.06$.

Пусть одна и та же ключевая последовательность использовалась несколько раз для разных сообщений одинаковой длины. При этом для всех сообщений, кроме последнего, известны открытые тексты. Шифртексты всех сообщений разделяются на s-граммы (с зацеплением) и посимвольно сопоставляются.

Введем случайную величину:

$$\xi_{i_j} = \begin{cases} 1, & \text{если для } i_j\text{-го символа} \\ \text{найдено хотя бы 1 совпадение} \\ 0, & \text{если совпадений нет.} \end{cases}$$

В местах совпадений неизвестный символ последнего сообщения определяется по информации об открытых символах предыдущих сообще-

ний. Тогда количество l_{i_j} вариантов знаков для всех отрезков неизвестного сообщения:

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

где m – мощность алфавита, M – число сообщений на одинаковом ключе.

Пусть вероятность $P\{\xi_{i_j} = 1\} = p$, $P\{\xi_{i_j} = 0\} = 1 - p$ соответственно для любых i и j.

Средней характеристикой i-го отрезка сообщения *в случае повторного шифрования* называется случайная величина:

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

Очевидно, что $\hat{S}_i$ – это количество символов на i-м отрезке, для которых удалось восстановить исходный знак.

Математическое ожидание и дисперсия величины $\hat{S}_i$:

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

Пусть значение критической границы L фиксировано. Тогда справедливы следующие утверждения о распределении вероятности появления L-ограниченного отрезка в неизвестном сообщении.

Вероятность того, что среднее геометрическое любого i-го отрезка неизвестного сообщения не превосходит заданной границы L, имеет биномиальное распределение с параметрами s и p:

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

для любого i, где C_s^k – биномиальный коэффициент, где $v = \left[s - s\frac{\log_2 L}{\log_2(m - M + 1)}\right]$.

Ковариация k-го и m-го произвольных отрезков сообщения длиной s символов равна:

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

Для сообщения длиной N символов рассмотрим биномиальный вектор:

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

где $M(s)$ – средняя характеристика i-й s-граммы сообщения, имеющая биномиальное распределение с параметрами s и p.

Вероятность того, что все отрезки сообщения длины s одновременно являются L-ограниченными, имеет мультиномиальное распределение с вектором средних значений θ и ковариационной матрицей Σ:

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

5.3. Математическая модель распределения числа осмысленных текстов

Пусть $N_s = m^s$ – общее количество всех s-грамм в алфавите мощностью m, среди которых присутствует ровно $M(s)$ осмысленных s-грамм. Величина $M(s)$ оценивается как $2^{H_s \cdot s}$, где H_s – энтропия s-грамм в расчете на один символ [12, 13].

Далее определим n_i как число возможных вариантов восстановления для i-го отрезка, среди которых присутствует один истинный. Поскольку состав выходных множеств для каждого неизвестного знака сообщения образуется случайным равновероятным выбором символов из исходного алфавита, мы рассматриваем множество $n_i - 1$ (ложных вариантов восстановления) как выборку на множестве $N_s - 1$, в котором $M(s) - 1$ осмысленных текстов.

Тогда вероятность, что в выборке из $n_i - 1$ различных вариантов восстановления ровно r_i окажутся осмысленными, описывается *гипергеометрическим распределением* [17], т.е.

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

Если восстанавливается i-й отрезок сообщения (мощность алфавита – m символов) длиной s символов со средним числом значений L_i, то вероятность появления r_i осмысленных вариантов восстановления для данного отрезка:

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

Данное утверждение следует из формулы (18). При этом вероятность найти ровно 1 осмысленный текст (истинный) оценивается при значении параметра $r_i = 0$.

Таким образом, вероятность успеха восстановления i-й s-граммы со средним числом значений L_i с учетом допустимой многозначности определяется как:

$$P_{\text{однозн}}(s, L_i, \beta) = \sum_{r_i=0}^{r_{\max}-1} P(L_i^s - 1, r_i), \qquad (20)$$

где $r_{\max} = \lfloor 2^{\beta s} \rfloor$ – максимально допустимая степень неоднозначности восстановления s-граммы, β – параметр алгоритма.

Предельное распределение числа осмысленных текстов возникает, когда $s \to \infty$. Тогда параметры $N_s = m^s \to \infty$, $M(s) = 2^{Hs} \to \infty$, H – предельное значение энтропии H_s при $s \to \infty$. При этом полагаем $L_i \leqslant \frac{m}{2}$. Вид предельного распределения зависит от количества осмысленных текстов в выборке.

Поскольку нижеприведенные утверждения 9–10 описывают случаи предельного распределения при $s \to \infty$, в них учитывается не энтропия коротких s-грамм, а соответствующее предельное значение энтропии (т.е. энтропия языка).

Вероятность найти ровно 1 осмысленный текст (истинный) при $s \to \infty$:

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

Используя формулу Стирлинга и второй замечательный предел:

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

$$= \frac{(N_s - M(s))!(N_s - n_i)!}{(N_s - M(s) - n_i + 1)!(N_s - 1)!} \to$$

$$\to \frac{(N_s - M(s))^{N_s - M(s)} \cdot (N_s - n_i)^{N_s - n_i}}{(N_s - M(s) - n_i + 1)^{N_s - M(s) - n_i + 1} \cdot (N_s - 1)^{N_s - 1}} \to$$

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

Вероятность получить $r_i = \lfloor 2^{\beta s} \rfloor$ осмысленных текстов, среди которых один истинный, при восстановлении i-го отрезка длиной $s \to \infty$ имеет вид:

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

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

При использовании утверждений 8-9 для численных расчетов значение параметра s должно удовлетворять условию $H_s \approx H$. Для русского языка $s \geqslant 50$, $H \approx 0.78$.

Если допустимое количество осмысленных текстов зависит от s, то $r_i = \lfloor 2^{\beta s} \rfloor \to \infty$. Предполагая параметры $\beta < 1$; $L_i \leqslant \frac{m}{2}$:

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

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

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

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

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

В разделе 8 приводятся некоторые численные расчеты доказанных утверждений.

6. ЭФФЕКТИВНОСТЬ АЛГОРИТМА

На практике эффективность алгоритма можно рассматривать как общую долю информации, которую удалось восстановить на выходе канала связи. При теоретических оценках эта доля определяется вероятностью восстановления L-ограниченного отрезка заданной длины, т.е. совместной вероятностью появления такого отрезка в сообщении и успеха его восстановления с учетом допустимой многозначности.

Тогда общая вероятность восстановления отрезков сообщения, среднее значение вариантов которых не превышает заданную критическую границу L, *без учета степени покрытия словаря* оценивается как:

$$P_{\text{восст}}(s, L, \beta) = P_{\text{появл}}(s, L) \cdot \sum_{L_i < L} P_{\text{однозн}}(s, L_i, \beta), \qquad (23)$$

где $P_{появл}(s, L)$ – вероятность появления i-го отрезка длины s со среднем числом вариантов, не превышающим L; $P_{однозн}(s, L_i, \beta)$ – вероятность, что восстановление i-го отрезка со среднем числом вариантов L_i не превысит допустимую многозначность.

С учетом *неполноты покрытия* используемого словаря общая доля восстановленной информации на выходе канала связи имеет вид:

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

где τ_s – степень покрытия словаря s-грамм.

Если общая доля восстановленной информации на выходе канала связи превышает установленное критическое значение $\pi > \pi_0$, то данный канал связи оценивается как *незащищенный*.

7. ТРУДОЕМКОСТЬ АЛГОРИТМА

Трудоемкие этапы составления и сортировки словарей являются предварительными и не входят в расчет вычислительной сложности самого алгоритма. Таким образом, трудоемкость алгоритма определяется сложностью реализации этапа восстановления L-ограниченного отрезка.

Пусть L – критическая граница отбора отрезков сообщения, s – длина восстанавливаемого отрезка сообщения, D_s – объем словаря s-грамм. Тогда количество вариантов восстановления L-ограниченного отрезка, проверяемых на присутствие в словаре, не превышает L^s.

Асимптотическая сложность восстановления L-ограниченного отрезка при реализации бинарного алгоритма поиска по словарю имеет вид:

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

Общее количество L-ограниченных отрезков в сообщении длиной N символов приведено в утверждении 5.

8. ЧИСЛЕННЫЕ ОЦЕНКИ ПАРАМЕТРОВ АЛГОРИТМА ДЛЯ РУССКОГО ЯЗЫКА

Рассмотрим алгоритм восстановления сообщения на русском языке с фиксированным алфавитом открытого текста – *35 символов* (32 буквы кириллического алфавита, пробел, точка и запятая).

При расчетах использованы оценки энтропии s-грамм русского языка [12].

Поскольку для отрезков длиной более 50 символов значение энтропии стабилизируется, за уровень предельной энтропии принято значение 0.78 бит на символ.

Исследуемые параметры алгоритма приведены в табл. 4.

Теоретически наиболее вероятное число осмысленных текстов (с учетом истинного) на русском языке, которые будут найдены при восстановлении отрезка сообщения, приведено в табл. 4.

Следовательно, для русского языка восстановление 10-грамм при среднем числе значений больше 8–9 символов практически бессмысленно. Для 15-грамм максимальная эффективность восстановления будет для значения L не больше 12–13 символов.

Таблица 2. Энтропия русского языка

Длина отрезка, s	10	15	20	25	более 50
Энтропия, H_s	2.49	1.80	1.41	1.16	0.78

Таблица 3. Параметры алгоритма

Мощность алфавита кодовых символов, m	35
Параметр β	0.1
Степень допустимой многозначности, $r_{\max}$	См. табл. 1
Длина восстанавливаемого отрезка, s	10–25
Критическая граница отбора отрезков, L	8–24
Энтропия s-грамм, H_s	См. табл. 2
Вероятности появления выходных значений символов, p_k	$p_k = \frac{1}{35}, \forall k = 1,\ldots,35$

С увеличением длины отрезка сообщения степень неоднозначности восстановления открытого текста уменьшается. Однако для участков текста большой длины процесс составления словарей поиска становится весьма затруднительным. Проведенное моделирование позволяет выбрать *оптимальный параметр алгоритма* – длину восстанавливаемого отрезка сообщения – на основе максимально эффективного соотношения между степенью неоднозначности восстановления открытого текста и степенью покрытия соответствующего словаря.

Очевидно, такой параметр – минимальная длина s-граммы, при которой вероятность восстановления отрезка приближается к единице для любого среднего числа вариантов знака. На основе вероятностей гипергеометрического распределения определяется длина такой минимальной s-граммы для русского языка (с алфавитом в 35 символов) – **16 символов**.

Для ряда случаев конкретного вида распределения вероятностей $p_k = P(l_{i_j} = k)$ (числа значений i_j-го входного символа) несложно получить численные оценки эффективности алгоритма.

Пусть, например, количество значений для всех отрезков распределено *независимо и случайно равновероятно* от 1 до m, т.е. принимает значения с одинаковыми вероятностями для любого $k = 1,2,\ldots,m$. Значение критической границы L фиксировано.

Таблица 4. Наиболее вероятное число осмысленных текстов

$L\|s$	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

Таблица 5. Теоретическая доля восстановленных отрезков

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

Теоретическая оценка эффективности алгоритма для русского языка (доля восстановленных отрезков в сообщении без учета степени покрытия словарей) при разных значениях параметров s и L, когда количество l_{i_j} значений для всех отрезков распределено равномерно от 1 до 35, приведена в табл. 5.

Таким образом, с ростом среднего числа значений неизвестных символов для отрезка на русском языке длиной не менее 16 символов вероятность его восстановления увеличивается с учетом допустимой степени неоднозначности.

9. ЭКСПЕРИМЕНТАЛЬНОЕ ИССЛЕДОВАНИЕ АЛГОРИТМА

В рамках подтверждения построенной модели проводится сравнение теоретических оценок эффективности алгоритма с экспериментальными результатами. Для осуществления эксперимента используются программная реализация[2] предлагаемого алгоритма и созданный корпус публицистических текстов русского языка объемом 1 миллион символов[3], подробно описанный в [12, 13].

На основе текстового корпуса программно[4] создаются словари s-грамм, используемые при восстановлении сообщения (табл. 6).

Последовательно рассматривается восстановление отрезков сообщения длиной 10, 15, 20 и 25 символов. Общая длина сообщения $N = 1000$ символов. Для разных значений критической границы L определяется относительная доля успешно восстановленных отрезков.

Кроме того, для проведения сравнения при тех же параметрах оценивается теоретическая эффективность алгоритма. При расчетах количество осмысленных текстов длиной s символов принимается равным объему соответствующего словаря s-грамм (возможная неполнота покрытия словарей не учитывается).

10. ЗАКЛЮЧЕНИЕ

Мы рассмотрели задачу восстановления сообщения, передаваемого в дискретном канале связи, когда на выходе имеется информация о возможных значениях исходного текста, возникшая в результате внесения нарушений в работу элементов системы защиты информации. В отличие от предыдущих подходов, мы предложили алгоритм восстановления отдельных s-грамм сообщения, когда полное восстановление по имеющейся информации невозможно. Предлагаемая процедура состоит из нескольких этапов, включая предварительное составление словарей s-грамм. Оценку эффективности предлагаемого алгоритма мы провели в рамках равновероятной модели, с помощью которой определили оптимальные параметры алгоритма для русского языка и критерий незащищенности канала связи. Теоретические оценки эффективности подтвердились экспериментальной проверкой построенной модели алгоритма.

В ходе исследования мы рассмотрели различные виды распределений числа возможных символов выходного сообщения. В итоге обнаружили, что с увеличением длины отрезка текста вероятность нахождения L-ограниченной s-граммы, потенциально пригодной для восстановления, приближается к нормальному распределению независимо от исходного числа значений. В случае повторного использования ключа данная вероятность описывается биномиальным распределением. Кроме того, мы обнаружили, что нормальное распределение позволяет получить хорошее приближение для исходных вероятностей уже при небольших значениях длины s-граммы.

[2] Результат интеллектуальной деятельности создан в Национальном исследовательском университете “Высшая школа экономики”. https://github.com/Nastasian/recovery

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

[4] Результат интеллектуальной деятельности создан в Национальном исследовательском университете “Высшая школа экономики”. https://github.com/Nastasian/entropy

Таблица 6. Словари s-грамм

Длина отрезка, s	10	15	20	25
Объем словаря, D_s	7.9×10^5	9.5×10^5	9.8×10^5	9.9×10^5
Энтропия s-грамм, H_s	1.96	1.32	0.99	0.80

Таблица 7. Оценка эффективности алгоритма

L	Экспериментальная оценка				Теоретическая оценка			
	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

Таблица 8. Оптимальные параметры алгоритма для русского языка

Длина восстанавливаемого отрезка сообщения $s = 16$ символов	Эффективность восстановления	
Максимальное число возможных значений для символа s-граммы	14 символов	50%
	16 символов	75%
	18 символов	90%

При исследовании степени неоднозначности восстановления текстовых сегментов, возникающей в результате применения предлагаемого алгоритма, мы использовали гипергеометрическое распределение. С помощью него получили численные расчеты вероятности появления допустимого количества осмысленных текстов при восстановлении одного отрезка. Кроме того, нашли асимптотику гипергеометрического распределения для больших значений длины сегмента.

Общую эффективность алгоритма рассмотрели как среднюю долю восстановленной информации в неизвестном сообщении. Для оценки сложности реализации алгоритма определили его трудоемкость, зависящую, в основном, от сложности перебора по словарю. В результате исследования эффективности восстановления для русского языка (без учета степени покрытия словаря) определили оптимальные параметры алгоритма (табл. 8).

Основное применение полученные результаты имеют в криптографии при проведении анализа стойкости криптографических алгоритмов, в частности, в ситуациях некорректного выбора ключа или его повторного использования в алгоритмах симметричного шифрования [21], а также в случаях различных утечек информации о символах исходного сообщения.

СПИСОК ЛИТЕРАТУРЫ

1. *Gorbenko I., Kuznetsov A., Lutsenko M., Ivanenko D.* The research of modern stream ciphers. 4th International Scientific-Practical Conference Problems of Infocommunications, Science and Technology (PIC S T), IEEE, 2017. P. 207–210.
2. *Aumasson J.P.* Serious cryptography: a practical introduction to modern encryption. No Starch Press, 2017.
3. *Rubinstein-Salzedo S.* Cryptography. Cham, Switzerland, Springer, 2018. P. 259.
4. *Mohamed Barakat*, Eder Ch. An introduction to cryptography. Technische Univ. Kaiserlautern, 2018. P. 138.
5. *Kessler G.C.* www.garykessler.net/library/crypto.html – An overview of cryptography, 2020.
6. *Гнеденко Б.В., Беляев Э.К., Соловьев А.Д.* Математические методы в теории надежности. Либроком, 2019. 584 с.
7. *Савинов А., Иванов В.* Анализ решения проблем возникновения ошибок первого и второго рода в системах распознавания клавиатурного почерка. Вестник ВУиТ. 2011. № 18.
8. *Babash A.V.* Attacks on the Random Gamming Code. Mathematics and Mathematical Modeling, 2020. № 6. P. 35–38.
9. *Бабаш А.В., Баранова Е.К., Лютина А.А., Мурзакова А.А., Мурзакова, Е.А., Рябова Д.М., Семис-Оол Е.С.* О границах зашумления текстов при сохранении их содержания. Приложения к криптографии. Вопросы кибербезопасности. 2020. № 1(35), С. 74–86.
10. *Coecke B., Fritz T., Spekkens R.W.* A mathematical theory of resources. Information and Computation, 2016. P. 59–86.
11. *Малашина А.Г., Лось А.Б.* Построение и анализ моделей русского языка в связи с исследованиями криптографических алгоритмов. Чебышевский сборник, 2022. V. 23. № 2. P. 151–160.
12. *Малашина А.Г.* Разработка инструментальных средств для исследования информационных ха-

рактеристик естественного языка. Промышленные АСУ и контроллеры. 2021. № 2. С. 9–15.

13. *Malashina A.* The combinatorial analysis of n-gram dictionaries, coverage and information entropy based on the Web Corpus of English. Baltic Journal of Modern Computing, 2021. V. 9. № 3. P. 363–376.
14. *Nurmukhamedov U.* Lexical coverage and profiling. Language Teaching 52, 2019. № 2. P. 188–200.
15. *Juzek T.S.* 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. № 158. P. 177–125.
16. *Morzy M., Kajdanowicz T., Kazienko P.* On measuring the complexity of networks: Kolmogorov complexity versus entropy. Complexity, 2017.
17. *Greene W.H.* Limited dependent variables – truncation, censoring and sample selection. Econometric analysis, 2008. P. 950.
18. *Золотарев В.М.* Современная теория суммирования случайных величин. Российский фонд фундаментальных исследований. 1998. № 96–01–01920, С. 256—258.
19. *Taha M.A., Sahib N.M., Hasan T.M.* Retina random number generator for stream cipher cryptography. International Journal of Computer Science and Mobile Computing, 2019. V. 8. № 9. P. 172–181.
20. *Poonam J., Brahmjit S.* RC4 encryption – a literature survey. Procedia Computer Science. 2015. V. 46. P. 697–705.
21. *Pachghare V.K.* Cryptography and information security. PHI Learning Pvt. Ltd., 2019. P. 520.
22. *Shannon C.E.* A mathematical theory of communication. The Bell system technical journal. 1948. V. 27 (3). P. 379–423.
23. *Айвазян С.А.* Прикладная статистика: Основы моделирования и первичная обработка. Справочное издание. М.: Финансы и статистика, 1983. 471 с.

ABOUT THE POSSIBILITY OF RECOVERING MESSAGE BASED SIDE ON INFORMATION ABOUT THE ORIGINAL CHARACTERS

A. G. Malashina[a]

[a] *HSE University, Moscow, Russian Federation*

Presented by Academician of the RAS A.L. Semenov

In order to ensure secure information exchange in communication channels, a preliminary study of the correctness of the operation of the relevant information protection systems is necessary. Despite the fact that the mathematical algorithms used in such systems are correct and theoretically provide the correct statistical properties of the output stream compared to the input, 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 sensor). 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