Бабаш А.В.

Атаки на шифры гаммирования использующие q-слабые ключи.

НИУ ВШЭ,

РЭУ им. Г.В. Плеханова

СТ1+К1=Ш1 СТ2+К2=Ш2 СТ1+К1 –(СТ2+К2)=Ш1-Ш2

СТ1-Ст2+(К1-К2)=Ш1-Ш2

Содержательные тексты

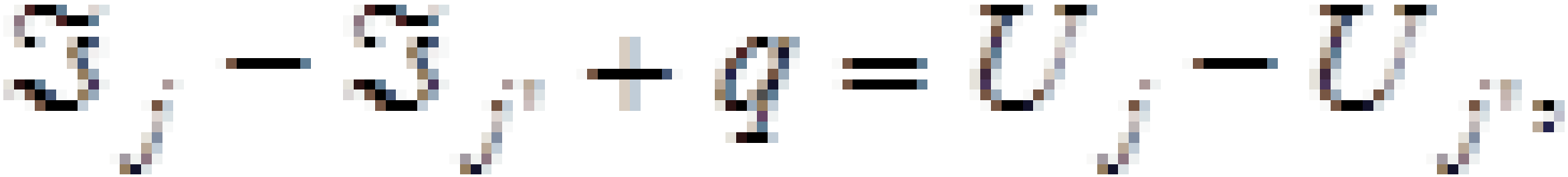
1100 0101 0010

ключ **1110**

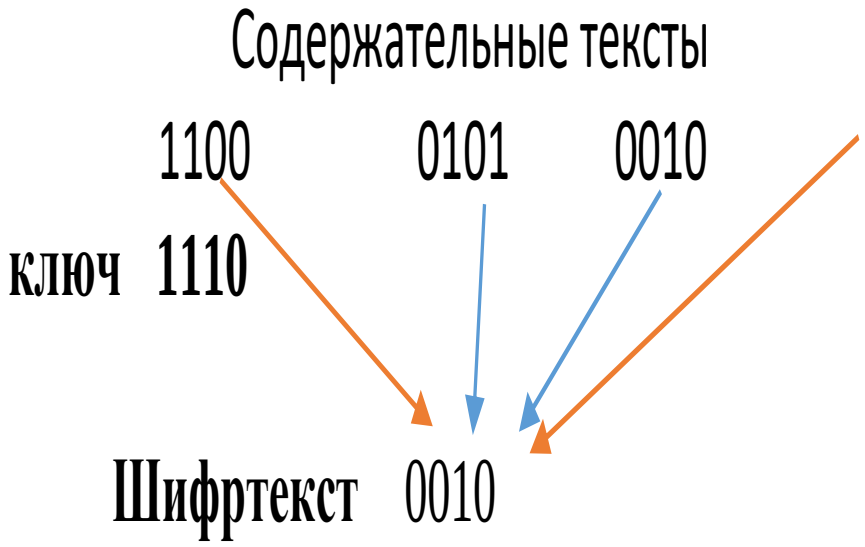
Шифр текст 0010

Позиции 2-грамм	1 2 разн		1 3 разн		2 3 раз	
Шифртекст **0010**	00-01=01		00-10=10		01-10=11	
СТ1= 1100	11-10=01		11-00=11		10-00=10	
Ключ 1110	11-11=**00**		11-10=01		11-10=01	
Шифртекст **0010**	00-01=01		00-10=10		01-10=11	
СТ2=0100	01-10=11		01-00=01		10-00=10	
Ключ 0110	01-11=10		01-10=11		11-10=01	
Выбранное q	00	01	10	11		
Результат №СТ	1	1,2	2	2		

- 2. Множество q-слабых ключей. Найдем q-сильные. D=2, L=4
- Перебор $(F_2)^2$ Взяли 01.
-
- 01000
- 01001
- 0101
- 01100
- 01101
- 0111
-



- M(L) – Содержательные тексты. M(q,U)- Найденные СТ.
- Красные стрелки K(q,U) –ключи соответствующие найденным M(q,U) СТ.



Всех ключей |K |, из них |K(q)| q-слабых.

- |M(q,U)|=| K(q,U)|=w   А всех СТ |M(L)|=|K(U)| - число стрелок- ключей.
- Случайная величина w распределена по гипергеометрическому распределению.

- Алгебраическая модель информации. Под информацией (криптографической) о содержательном истинном тексте (СТ) полученном при дешифровании шифра понимаю не пустое подмножество множество СТ содержащее истинный текст. Под ценой информации понимаю дополнение к информации. Под дешифрованием понимаю определение информации о зашифрованном СТ.
- Вероятностная модель информации. Как у К. Шеннона на множестве содержательных текстов задано вероятностное распределение. Информацией о содержательном тексте называется фрагмент вероятностного распределения на M(L), содержащий истинный текст. Ценой информации называется дополнение к информации. Дешифрованием называется определение информации о зашифрованном СТ.

1.Введение. Ранее были опубликованы 8 алгоритмов дешифрования шифра случайного гаммирования (ШСГ)с расчетом их трудоемкости и надежности. Основные идеи дешифрования были основаны на наличии ключей шифра Виженера в качестве подмножества ключей шифра случайного гаммирования и известных к настоящему времени методов дешифрования последнего шифра. В докладов автора на конференциях РусКрипто 19, РусКрипто 21 и МиТСОБИ 20 был разработан новый метод дешифрования шифра случайного гаммирования, целью которого явилось получение информации о D-граммах передаваемого содержательного текста. В данном докладе излагается метод определения информации о зашифрованном содержательном тексте ШСГ. Вычисляются параметры сложности указанного метода. Приводятся примеры. Результаты переносятся на шифры гаммирования.

Модель шифра случайного гаммирования. Пусть I, K, Y - алфавиты используемого языка, ключей, шифрованных текстов. Положим $I = K = Y = Z/s$ - кольцо положительных вычетов по модулю s. $M(L)$ - известное *перечислимое множество содержательных (СТ) текстов длины L подлежащих шифрованию, элементы которых пронумерованы*. Через $i + \gamma = y \bmod s$ обозначим функцию шифрования ШСГ для букв $i \in I, \gamma \in K, y \in Y$. Уравнение расшифрования имеет вид $y - \gamma = i \bmod s$. Для шифрования открытого текста $\mathfrak{I}_j^\nu = i_j^\nu i_{j+1}^\nu \ldots i_{j+d-1}^\nu$ из I^d выбирается случайно и равновероятно ключ $\Gamma_j^\nu = \gamma_j^\nu \gamma_{j+1}^\nu \ldots \gamma_{j+d-1}^\nu$ проводится операция шифрования $\mathfrak{I}_j^\nu + \Gamma_j^\nu = U_j^\nu = y_j^\nu y_{j+1}^\nu \ldots y_{j+d-1}^\nu$, где $i_{j+c}^\nu + \gamma_{j+c}^\nu = y_{j+c}^\nu \bmod s$, $c \in \{0,1,\ldots d-1\}$.

2.1 Обозначение операции обратного элемента кольца перенесем и на $(Z/s)^d$. Уравнение расшифрования имеет вид $U_j^\nu - \Gamma_j^\nu = \mathfrak{I}_j^\nu$. Шифрование содержательного текста $\mathfrak{I}^\nu = \mathfrak{I}_1^\nu \mathfrak{I}_2^\nu \ldots \mathfrak{I}_n^\nu$, $n = \frac{L}{d}$ и расшифрование шифртекста $U^\nu = U_1^\nu U_2^\nu \ldots U_n^\nu$ с помощью ключа $\Gamma^\nu = \Gamma_1^\nu \Gamma_2^\nu \ldots \Gamma_n^\nu$ проводится по приведенным выше законам шифрования и расшифрования d-грамм. Ключи Γ для шифрования текстов из I^L выбираются случайно и равновероятно из K^L и независимо от текста из I^L. Решаемая задача состоит в определении информации о передаваемом содержательном тексте по известному шифрованному тексту шифра случайного гаммирования.

3.Определение q-слабого ключа. Используемый ключ $\Gamma^{\nu} = \Gamma_1^{\nu}\Gamma_2^{\nu}...\Gamma_n^{\nu}$ назовем q-слабым ключом, если найдется пара позиций j, j', $j < j'$ удовлетворяющих *условию* $\Gamma_j - \Gamma_{j'} = q.$ Множество всех q-слабых ключей обозначим через $K_q(d,L) \subset K^L$, а вероятность использования q-слабого ключа обозначим через $P_q(d,L) = \frac{K_q(d,L)|}{K^L}$.

4 ***Примеры поиска информации*** $M(q,U)$. Напомним уравнения

$$\Im_j + \Gamma_j = U_j, \quad \Im_{j'} + \Gamma_{j'} = U_{j'}$$

шифрования для двух d-грамм j, j' содержательного текста $\Im = \Im_1\Im_2...\Im_n$, длины L=nd и следствие из них $\Im_j - \Im_{j'} + q = U_j - U_{j'}$(*), где $\Gamma_j - \Gamma_{j'} = q.$ При известном шифрованном тексте $U = U_1U_1...U_n$ и зафиксированном q конструктивно определены множества $M_q(j, j', U)$ всех содержательных текстов $\Im = \Im_1\Im_2...\Im_n$, зашифрованных в U, у которых d-граммы $\Im_j, \Im_{j'}$ начинающиеся с позиций j, j', $j < j'$, удовлетворяют *условию* $\Im_j - \Im_{j'} + q = U_j - U_{j'}$ Множество $M_q(U) = \bigcup_{j,j'} M_q(j, j', U)$ есть множество СТ зашифрованных q-слабыми ключами.

5. Пример 1. s=3, q=0, d=2, L=4, $M(L)=\{CT1=2220, CT2=1210\}$. Шифртекст 0101. Расчеты представлены таблицей 1

Таблица 1

Позиции 2-грамм	1 - 2	разность	1 - 3	разность	2 - 3	разность
Шифртекст **0101**	01 - 10	21	01 - 01	00	10 - 01	12
CT1= 2220	22 - 22	00	22 - 20	02	22 - 20	02
Ключ 1211 для 2220	12 - 21	**21**	12 - 11	01	21 - 11	10
Шифртекст **0101**	01 10	21	01 - 01	00	10 - 01	12
CT2= 1210	12 21	21	12 - 10	02	21 - 10	11
Ключ 2221 для 1210	22 22	**00**	22 - 21	01	22 - 21	01
Выбранное q	00	01	21	01	10	
Результат №CT	2	2	1	1	1	

Таблица 1. Расчеты примера 1.

Результатом q=0–атаки явилась информация {CT2= 1210} при q=0-слабым ключом. И {CT1} для q=21 (см. таблицу 1).

6.Пример 2. s=3, q=0, d=2, L=4, $M(L)=\{CT1=2021, CT2=1210\}$, шифртекст 0101. Расчеты представлены таблицей 2

Таблица 2

Позиции	j_1, j_2	1 2 разность	1 3 разность	2 3 разность
Шифртекст	0101	01 10 **21**	01 01 00	10 01 **12**
СТ1= 2021		20 02 **21**	20 21 02	*02 21* ***11***
Ключ 1110 для	2021	11 11 **00**	11 10 01	11-10 **01**
Шифртекст	0101	01 10 **21**	01 01 00	10 01 **12**
СТ2= 1210		12 21 21	12 10 02	21 10 11
Ключ 2221 для	1210	22 22 00	22 21 01	22-21= 01

Таблица 2. Расчеты примера 2.

Результатом q=0–атаки является информация – множество текстов $M(q=0,U)=M(L)=\{2021,1210\}$ зашифрованных q=0-слабыми ключами. То есть новой информации о зашифрованном тексте не получено.

7. Пример 3. s=3, q=0, d=2, L=4, $M(L)=\{1020,2220,1210\}$ шифртекст 0101. Расчеты представлены таблицей 3

Таблица 3

Позиции j_1, j_2	1 2 разность	1 3 разность	2 3 Разность
Шифртекст 0101	01 10 21	01 01 00	10 01 12
СТ 1020	10 02 11	10 20 20	02 20 12
СТ 2220	22 22 00	22 20 02	22 20 02
СТ 1210	**12** **21** 21	12 - 10 02	21 10 11

Таблица 3. Расчеты примера 3.

Истинный передаваемый СТ содержится в множестве $M(q=0,U)=\{1020,1210\}$. Для q=21 найден СТ=2220.

8. Полное определение содержательного текста можно получить из полученной информации $M(q,U)$ путем выбора наиболее вероятного варианта. В этом случае расчет вероятности правильного определения содержательного текста очевиден, как и процесс переноса алгоритма q-дешифрования на произвольные шифры гаммирования с моделью ключа –гаммы шифрования.

9. **Алгоритм подсчета вероятности** $P_{q=0}(d,L)$.

Обозначим через $W(\gamma_1=0,d,L)$ множество всех ключей $\Gamma=0\gamma_2...\gamma_L$ длины L без повторений d-грамм (q=0…0) с начальным символом $\gamma_1=0$. Тогда

$$K_{q=0}(d,L)=(Z/s)^L \setminus W(\gamma_1=0,d,L).$$

Шаг 1. Строим множество $W(\gamma_1=0,d,L=d+1)$ всех ключей $\Gamma=0\gamma_2...\gamma_{d+1}$ длины $L=d+1$ без повторений d-грамм (q=0…0) с начальным символом $\gamma_1=0$. С этой целью проводим перебор элементов $0\gamma_2...\gamma_d$ из $W(\gamma_1=0,d,L=d)$. Для каждого слова $0\gamma_2...\gamma_d$ находим путем сравнения слов все элементы $\gamma\in Z/s,$ при которых нет повторений d-грамм в слове $0\gamma_2...\gamma_d\gamma$. Здесь возможны повторения лишь вида $0\gamma_2...\gamma_d=\gamma_2...\gamma_d\gamma$. В результате будет найдено множество $W(\gamma_1=0,d,L=d+1)$ слов длины d+1 с начальным символом 0 без повторений d-грамм. Трудоемкость шага 1 оцениваем в s^{d+1} операций.

10. Продолжение алгоритма.

Практически проведен индукционный шаг. Следовательно, построены множества $W(\gamma_1 = 0, d, L = d + c)$, $c \in \{1, ...L - d\}$. Очевидно можно считать, что $L - d \le s^d - 1\}$. Общее число операций будет равно

$$T(L) = |W(\gamma_1 = 0, d, d)| \cdot s \cdot 1 + |W(\gamma_1 = 0, d, d + 1)| \cdot s \cdot 2 + ... + |W(\gamma_1 = 0, d, L - 1)| \cdot s \cdot (L - D)$$

В числе сравнений слов на шаге построения множества $|W(\gamma_1 = 0, d, d + j + 1)|$ учитывались сравнения только последней построенной d-граммы с d-граммами множества $|W(\gamma_1 = 0, d, d + j)|$. Их число равно $j + 1$.

В результате найдено множество $W(\gamma_1 = 0, d, L)$ для $L \ge d$. Заметим, что любые два слова отличающиеся в каждой позиции на фиксированный элемент из Z / s содержат одинаковое число повторяющихся d-грамм. Поэтому число ключей с неповторяющимися d-граммами равно $s|W(\gamma_1 = 0, d, L)|$ при $L \ge d$. Число $s^L - s \cdot |W(\gamma_1 = 0, d, L)|$ представляет собой число ключей с повторяющейся хотя бы одной d-граммой. Таким образом определена вероятность $P_{q=0}(d, L) = 1 - \frac{|W(\gamma_1 = 0, d, L,)|}{s^{L-1}}$.

11. Алгоритм подсчета q-слабых ключей в случае q>0. Подсчет вероятности $P_q(d,L) = \frac{K_q(d,L)|}{K^L}$. в случае $q \neq 0$ проводится аналогично с помощью перечисления множества $K_q(d,L)$. При этом следует учесть, что свойство симметричности по первой компоненте γ_1 ключа не выполняется. Алгоритм проводится для каждой из s возможных первых компонент ключа с последующим суммированием результатов.

Алгоритм подсчета q(w)-слабых ключей. Данный алгоритм перечисляет q-слабые ключи с ровно w повторениями пар d-грамм с разностью q. Идейная и техническая сторона алгоритма остаются прежними. К опробуемому отрезку ключа добавляются пометка о числе w’<w предыдущих пар d-грамм с разностью q.

12. **Описание алгоритма q-дешифрования**. Фиксируем $q \in (Z/s)^d$ и шифрованный текст $U = U_1 U_2 ... U_n$. Обозначим через $M_q(j, j', U)$ множество всех содержательных текстов $\mathfrak{I} = \mathfrak{I}_1 \mathfrak{I}_2 ... \mathfrak{I}_n$, зашифрованных в U, у которых d-граммы $\mathfrak{I}_j, \mathfrak{I}_{j'}$ начинающиеся с позиций $j, j', \ j < j'$, удовлетворяют *условию*

$$\mathfrak{I}_j - \mathfrak{I}_{j'} + q = U_j - U_{j'}, \qquad (2)$$

Параметрами сложности процедуры определения множеств $M_q(j, j', U)$ являются $|M(L)|$ и число пар позиций $\frac{(\mathrm{L}-d+1)(\mathrm{L}-d)}{2}$, для которых проверяется выполнение равенства (2). Поэтому трудоемкость определения множества $M_q(U) = \bigcup_{j,j'} M_q(j, j', U)$ равна $\frac{(\mathrm{L}-d+1)(\mathrm{L}-d)}{2} |M(L)|$. Множество $M(L)$ представимо в виде разбиения

$$M(L) = M_q(U) \bigcup (M(L) \setminus M_q(U)). \qquad (3)$$

Множество $M_q(U)$ является множеством всех СТ из $M(L)$ зашифрованных в U на ключах из подмножества $K_{U,q}(d, L)$ множества $K_q(d, L)$. Поэтому $|M_q(U)| = |K_{U,q}(d, L)|$. И задача определения $|M_q(U)|$ равносильна определению мощности множества $K_{U,q}(d, L)$.

13. **Расчет множества** $K_{U,q}(d,L)$. Множество $M_q(U)$ является множеством всех СТ из $M(L)$ зашифрованных в U на ключах из подмножества $K_{U,q}(d,L)$ множества $K_q(d,L)$. Именно для $\Im \in M_q(U)$ ключ $\Gamma \in K_U(q,d,L)$ является решением уравнения $\Im + \Gamma = U$ когда он q-слабый. Поэтому $|M_q(U)| = |K_{U,q}(d,L)|$. И задача определения $|M_q(U)|$ равносильна определению мощности множества $K_{U,q}(d,L)$. Для ее решения естественно предположить, что множество $K_{U,q}(d,L)$ выбирается из множества $K_q(d,L)$ случайно и равновероятно. Таким образом вероятностное распределение случайной величины $|K_{U,q}(d,L)|$ (числа успехов) является гипергеометрическим распределением. В нашем случае гипергеометрическое распределение описывает вероятность $P(|K_{U,q}(d,L)| = w, s^L, |K_q(d,L)|, |M(L)|)$ того, что в выборке из $|M(L)|$ различных ключей ровно w ключей являются q-слабыми, при этом всего ключей s^L, из них $K_q(d,L)$ ключей q-слабые.

$$P(w, s^L, |K_q(d,L)|, |M(L)|) = \frac{\binom{|K_q(d,L)|}{w}\binom{s^L - |K_q(d,L)|}{|M(L)| - w}}{\binom{s^L}{|M(L)|}},$$ где w лежит в промежутке T

$$T = \max\{0, |K_q(d,L)| + |M(L)| - s^L\} \text{ и } \min\{|M(L)|, |K_q(d,L)|\}.$$

Среднее значение случайной величины $K_{U,q}(d,L) = w$ равно

$$|E(K_{U,q}(d,L)|) = \frac{|M(L)||K_q(d,L)|}{s^L} = |M(L)|P_q(d,L)$$

14. Известно, что в случае когда число ключей s^L намного больше чем число содержательных чисел $|M(L)|$, гипергеометрическое распределение хорошо аппроксимируется биномиальным распределением с параметрами $|M(L)|$ (количество испытаний) и $P_q(d,L)$ (вероятность успеха в одном испытании).

16. Понятия: информация о содержательном тексте, дешифрование шифра.

Мы с Вами понимаем термин «дешифрование» по-разному. Спорить о терминах можно долго и бесполезно.

Вот как его понимают Погорелов Б.А. и Сачков В.Н. «Дешифрование [decryption, breaking of cryptosystem] — процесс аналитического раскрытия противником и/или нарушителем сообщения открытого без предварительного полного знания всех элементов системы криптографической. Если этот процесс поддается математической формализации, говорят об алгоритме дешифрования».

17. Наверно Вы согласитесь, что дешифруемость корректно определять, как отрицание недешифруемости. Вот что говорят классики криптографии.

К. Шеннон - понятие недешифруемости: «естественно определить *совершенную секретность* с помощью следующего условия: для всех криптограмм апостериорные вероятности равны априорным вероятностям независимо от величины этих последних. В этом случае перехват сообщения не дает шифровальщику противника никакой информации. Теперь он не может корректировать никакие свои действия в зависимости от информации, содержащейся в криптограмме, так как все вероятности, относящиеся к содержанию криптограммы, не изменяются. С другой стороны, если это условие равенства вероятностей не выполнено, то имеются такие случаи, в которых для определенного ключа определенных выборов сообщений апостериорные вероятности противника отличаются от априорных. А это в свою очередь может повлиять на выбор противником своих действий и, таким образом, совершенной секретности не получится».

18 *А.П. Алферов, А.Ю. Зубов, А.С. Кузьмин, А.В. Черемушкин* «При рассмотрении вопроса о теоретической стойкости шифров отвлекаются от реальных временных и сложностных затрат по вскрытию шифра (что определяет подход к практической стойкости). Во главу угла ставится принципиальная возможность получения некоторой информации об открытом тексте или использованном ключе. Впервые такой подход исследовал К. Шеннон]. Он рассматривал уже знакомую нам модель шифра и единственную криптоатаку на основе шифртекста. Проследим за его рассуждениями. Как мы указывали, конечной целью работы криптоаналитика является текст сообщения или ключ шифрования. Однако весьма полезной может быть даже некоторая вероятностная информация об открытом тексте. Например, уже предположение о том, что открытый текст написан по-английски, представляет криптоаналитику определенную априорную информацию об этом сообщении даже до того, как он увидит шифртекст»

19. *Bruce Schneier*. «The hotline between the United States and the former Soviet Union was (is it still active?) rumored to be encrypted with a one-time pad. Many Soviet spy messages to agents were encrypted using one-time pads. These messages are still secure today and will remain that way forever. It doesn’t matter how long the supercomputers work on the problem. Even after the aliens from Andromeda land with their massive spaceships and undreamed-of computing power, they will not be able to read the Soviet spy messages encrypted with one-time pads (unless they can also go back in time and get the one-time pads»

20.Таких однотипных определений я могу привести более 10.

А что может и нет альтернатив высказываниям уважаемых криптографов?

Шерлок Холмс – «Любая тайна, порожденная человеческим сознанием, им же может быть и раскрыта»

Эдгар По – «…едва ли разуму человека дано загадать такую загадку, которую разум другого его собрата, направленный должным образом, не смог бы раскрыть.

Лаврентий Павлович Берия – «Нет не дешифруемых шифров».

21. *Теперь мое понимание*.

Алгебраическая модель информации. Под информацией (криптографической) о содержательном истинном тексте (СТ) полученном при дешифровании шифра понимаю не пустое подмножество множество СТ содержащее истинный текст. Под ценой информации понимаю дополнение к информации. Под дешифрованием понимаю определение информации о зашифрованном СТ.

Вероятностная модель информации. Как у К. Шеннона на множестве содержательных текстов $M(L)$ задано вероятностное распределение. Информацией о содержательном тексте называется не пустое подмножество элементов (с их вероятностями) множества $M(L)$ содержащее истинный текст. Ценой информации называется дополнение к информации. Дешифрованием называется определение информации о зашифрованном СТ.

22. Если Вы умеете по шифрованному тексту заданного шифра определять каждый 10 знак переданного содержательного текста (СТ), то этот шифр снимают с линии связи как дешифруемый. Потому, что не дешифруемый шифр таким свойством не обладает. Считают, что шифр простой замены дешифруем. Но известно, что на коротких длинах шифртекста невозможно определить СТ. При длинах передаваемых текстов равных 1 как и в шифре случайного гаммирования (ШСГ), так и в шифре простой замены невозможно определить СТ длины 1. То есть понятие дешифруемости зависит не только от шифра, но и выбора его модели (параметров). Именно такое положение мы с Вами наблюдаем и с ШСГ. В изложенных атаках модель ШСГ состоит из фиксации, в частности, параметров d, L, $M(L)$, и характеристик $M(L)$ - результатов разностей d-грамм $M(L)$. Каждая из представленных q-слабых атак может быть, как удачной, так и неудачной.

23. Промежуточные выводы. Указанная выше модель $M(L)$ позволяет для каждого шифртекста U указать по критерию $\Im_j - \Im_{j'} + q = U_j - U_{j'}$ данное разбиение $M(L) = M_q(U) \bigcup (M(L) \setminus M_q(U))$. Следовательно, если отправитель зашифровал сообщение на q-слабом ключе, то по шифрованному тексту *определяется информация о сообщении - заведомо не пустое множество* $M(q,U)$ *СТ* содержащее отправленное сообщение. Вероятность этого события равна вероятности $P(q,d,L)$ использования q-слабого ключа.

Первое множество $M(q,U)$ из разбиения $M(L) = M_q(U) \bigcup (M(L) \setminus M_q(U)))$ назовем *информацией* о зашифрованном содержательном тексте, второе – *ценой* информации. При фиксации модели $M(L)$ и процессов шифрования ШСГ, например, фиксирования модели шифра К. Шеннона шифра можно оценить и другие вероятностные параметры q-атаки. Так, величина $|M(L)|P(q,d,L)$ есть средняя мощность информации о зашифрованном СТ, полученной q-атакой по шифрованному тексту *в предположении*, что множество ключей, переводящих $M(L)$ в случайно и равновероятно выбранный шифртекст U выбрано случайно и равновероятно.

24. Моделирование множества пар содержательных текстов с помощью их d-грамм с приложением к дешифрованию ШСГ.

Фиксируем множество M(L). Будем говорить, что имеется $\Im(d, j)$-грамма естественного языка, если она является d-граммой $i_j i_{j+1} \ldots i_{j+d-1}$ СТ $\Im = i_1 i_2 \ldots i_j i_{j+1} \ldots i_{j+d-1} \ldots i_L$ стоящей в тексте на j-том месте. Парной d-граммой $\Im(d, j_1 j_2)$ содержательного текста $\Im = i_1 i_2 \ldots i_L$ назовем две его $\Im(d, j_1)$-грамм, $\Im(d, j_2)$-грамм $\Im_{j_1} = i_{j_1} i_{j_1+1} \ldots i_{j_1+d-1}$ и $\Im_{j_2} = i_{j_2} i_{j_2+1} \ldots i_{j_2+d-1}$ для $1 \le j_1 < j_2 \le j_{L-d+1}$. Ниже в случае, когда текст $\Im$ в парной d-грамме $\Im(d, j_1 j_2)$ не указывается то есть $(d, j_1 j_2)$ будем считать, что существует в M(L) содержательный текст $\Im'$, при котором $(d, j_1 j_2) = \Im'(d, j_1 j_2)$. Через $M(d, j_1 j_2)$ обозначим множество всех парных d-грамм, положим $МП(d) = \bigcup_{j_1 j_2} M(d, j_1 j_2)$. Моделирование множества введенных параметров *при больших D* диктуется использованием q-слабых ключей в методах дешифрования ШСГ по шифртексту.

25. Модели естественного языка на основе d-грамм содержательных текстов. Очевидно в первом приближении моделью обычного естественного языка можно считать последовательное выписывание букв (буква за буквой) так, чтобы читаемость сохранялась. Этот процесс можно назвать использованием последовательной конкатенации букв согласно орфографии языка. Ниже мы копируем этот процесс в упрощенной форме для $(d, j_1 j_2)$-грамм при построении множеств $M\Pi(d), M(d, j_1 j_2)$.

26. *Модель* 1. Положим D=Vd, $2D \le L$. Будем предполагать, что правилами грамматики языка определен критерий читаемости и не читаемости двух *подряд идущих* d-грамм $i_1 i_2 ... i_d i_1' i_2' ... i_d'$, то есть *конкатенации* двух d-грамм. Этот же критерий мы будем использовать в качестве определения конкатенаций парных (КП) d-грамм из множества $МП(d)$. Формальное определение будет дано позднее. Далее будем считать, что если определены множества $МП(vd), M(vd, j_1 j_2), v \in \{1,2,...V-1\}$, то множества $МП((v+1)d), M((v+1)d, j_1 j_2), v \in \{1,2,...V-1\}$ определяются условием парной конкатенации последних двух парных d-грамм в парах (v+1)d-грамм. Таким образом, нами построена модель 1 множеств $МП(Vd), M(Vd, j_1 j_2)$.всех СТ длины Vd.

Модель 2. Положим $2D \le L, D > d$. Определим на множестве парных d-грамм $МП(d)$ другую конкатенацию – *плотную* конкатенацию *парных* d-грамм (ПКП). Пара jd- грамм $\Im_{j_1} = i_{j_1} i_{j_1+1} ... i_{j_1+d-1}$ и $\Im_{j_2} = i_{j_2} i_{j_2+1} ... i_{j_2+d-1}$ будет находится в отношении ПКП с парой jd-грамм $\Im'_{j_3}$ и $\Im_{j_4}$ если последние имеют вид $\Im'_{j_3} = i_{j_1+2} ... i_{j_1+d-1} i_{j_1+d}$, $\Im_{j_4} = i_{j_2+1} ... i_{j_2+d-1} i_{j_2+d}$, где $i_{j_1+d} \in I, i_{j_2+d} \in I$. Указанный индукционный шаг позволяет строить модель множества всех парных d+c-грамм, $d+c=D$.

27. Использование моделей 1,2 содержательных текстов в оценке мощности получаемой информации методом q-слабых ключей при дешифровании ШСГ.

Используем введенные ранее обозначения. Пусть заданы пары позиций $\begin{matrix} j_1 & j_{1+d} & \dots & j_{1+vd} & \dots & j_{1+(V-1)d} \\ j_2 & j_{2+d} & \dots & j_{2+vd} & \dots & j_{2+(V-1)d} \end{matrix}$. Для каждой пары позиций $\begin{matrix} j_{1+vd} \\ j_{2+vd} \end{matrix}$ известно *упорядоченное* подмножество пар d-грамм вида $M\begin{pmatrix} j_{1+vd} \\ j_{2+vd} \end{pmatrix} = \begin{Bmatrix} \Im^{t(v)}_{1+vd} \\ \Im^{t(v)}_{2+vd} \end{Bmatrix}$ множества M(Od). Причем

Из дальнейшего будет ясно, что без ограничения общности можно считать, что мощности множеств $M\begin{pmatrix} j_{1+vd} \\ j_{2+vd} \end{pmatrix}$ одинаковы. Обозначим эту мощность через h. Итак имеем последовательность множеств

$$\begin{Bmatrix} \Im^{t(1)}_{1+d} \\ \Im^{t(1)}_{2+d} \end{Bmatrix} \begin{Bmatrix} \Im^{t(2)}_{1+2d} \\ \Im^{t(2)}_{2+2d} \end{Bmatrix} \dots \begin{Bmatrix} \Im^{t(V-1)}_{1+(V-1)d} \\ \Im^{t(V-1)}_{2+(V-1)d} \end{Bmatrix}. \qquad (*)$$

28. *Для модели 1* задача состоит в оценке мощности η множества всех пар СТ длины Vd полученных с помощью *конкатенаций парных* (КП) соседних d-грамм в колонках. *Решение для модели 1*. Поясним сначала выбор пар последовательностей из последовательности пар упорядоченных множеств (*). Первым элементом строящейся последовательности (выборки) может быть любой элемент из первого множества, любой элемент из второго множества и так далее. Всех различных последовательностей h^V. **Из них не должна учитываться одна последовательность – пара истинных содержательных текстов**. Указанный процесс получения пар содержательных текстов назовем *зигзагообразным чтением* пар текстов. Из h^V-1 последовательностей пар d-грамм отбираются последовательности, в которых соседние пары d-грамм (КП), говоря проще это «читаемые» пары текстов длины Vd.

29. Для получения оценки числа η читаемых текстов решим сначала следующую *вспомогательную задачу*.

Задача. Дано: Язык в алфавите I. Число всех последовательностей длины C букв алфавита I равно $|I|^C$, |M(C)| – число читаемых текстов длины $C \in \{1,2,...\}$. Имеются h^C последовательностей, полученных выборкой из равновероятного распределения на множестве всех последовательностей длины C букв алфавита. Какова вероятность $P_C(m)$ того, что будет выбрано m читаемых текстов?

30. Первый этап адаптации задачи к модели 1 языка. Имеется всего h выборок, которые мы представляем последовательно строками букв длины C. Следовательно, имеем C колонок букв глубины h. Проводим построение h^C последовательностей букв выбирая из каждого j столбца каждую букву и приписывая справа к ней каждую букву следующей j+1 колонки. Вероятность $P_C(m)$ рассчитываем для множества из h^C текстов трактуемых как h^C последовательностей, полученных выборкой из равновероятного распределения. Дополнительно предполагаем, что в этом множестве заведомо содержится фиксированный нами читаемый текст. Поэтому число выборок будет равным h^C-1. Очевидно, что случайная величина m распределена по гипергеометрическому распределению

$$P_C(m) = \frac{\binom{|M(C)|-1}{m}\binom{|I|^C-|M(C)|}{h^C-1-m}}{\binom{|I|^C-1}{h^C-1}} \approx \frac{\binom{|M(C)|}{m}\binom{|I|^C-|M(C)|}{h^C-m}}{\binom{|I|^C}{h^C}}$$

Среднее значение m есть $E_C(m) = \frac{(M(C)-1)(h^C-1)}{|I|^C-1}$.

31. Подсчитаем асимптотику вероятности $P_C(0)$ того, что при зигзагообразном чтении будет найден только один *фиксированный нами читаемый текст*.

$$P_C(0) = \frac{\binom{|I|^C - |M(C)|}{h^C}}{\binom{|I|^C}{h^C}} = \frac{(|I|^C - |M(C)|)!(|I|^C - h^C)!}{(|I|^C - |M(C)| - h^C)!|I|^C!}.$$

32. Предположим, что длина сообщения $C \to \infty$. Очевидно, что при этом $|M(C)| \to \infty$, $h^C \to \infty$ и $\frac{|M(C)|}{|I|^C} = o(1), \frac{h^C}{|I|^C} = o(1)$. С использованием формулы Стирлинга $n! = \sqrt{2\pi n} \cdot n^n e^{-n}(1 + o(1))$ получаем

$$P_C(0) = \frac{\left(1 - \frac{|M(C)|}{|I|^C}\right)^{|I|^C - |M(C)|} \left(1 - \frac{h^C}{|I|^C}\right)^{|I|^C - h^C}}{\left(1 - \frac{|M(C)| + h^C}{|I|^C}\right)^{|I|^C - (|M(C)| + h^C)}} (1 + o(1)) \approx$$

$$\approx \exp\left\{ -\frac{(|I|^C - |M(C)|)|M(C)|}{|I|^C} - \frac{(|I|^C - h^C)h^C}{|I|^C} + \frac{(|I|^C - (|M(C)| + h^C))(|M(C)| + h^C)}{|I|^C} \right\} =$$

$$= \exp\left\{ -2\frac{|M(C)| \cdot h^C}{|I|^C} \right\}.$$

33. Аппроксимируем рост мощности $|M(C)|$, положив $|M(C)| \approx |I|^{H_C \cdot C}$, где H_C=constant. Тогда $P_C(0) \approx \exp\left\{-2\frac{h^C}{|I|^{C-H_C \cdot C}}\right\} = \exp -2(\frac{h}{|I|^{1-H_C}})^C$.

Отсюда следует, что если $h < |I|^{1-H_C}$, то *вероятность единственности зигзагообразного чтения* стремиться к единице, а при $h > |I|^{1-H_C}$ – к нулю. Величину критической глубины h колонок (зигзагообразного чтения) можно положить равной $h = |I|^{1-H_C}$.

34. Второй этап адаптации задачи к модели 1 языка. Дано: искусственный язык в алфавите I. Предположим, что два человека *одновременно* используют (говорят, пишут) на естественным языке в алфавите I. Теперь формально. Алфавит Ω искусственного языка есть прямое произведение $\Omega = I^d \times I^d = \{ \begin{matrix} i_1 i_2 ... i_d \\ i_1' i_2' ... i_d' \end{matrix} : i \in I, i' \in I \}$ алфавитов I^d на I^d.

Число всех последовательностей длины V букв алфавита Ω равно $|\Omega|^V$. Читаемыми текстами в алфавите Ω назовем случай последовательности конкатенируемых d-грамм алфавита Ω - двух последовательностей в алфавите I, каждая из которых читаема в алфавите I. Пусть как и ранее |M(dV)| – число читаемых текстов длины dV в алфавите I. Тогда за число читаемых текстов в алфавите Ω длины V можно принять $|M(V)|^2$.

35. *Задача:* имеются h^V последовательностей, полученных выборкой из равновероятного распределения на множестве всех последовательностей длины V букв алфавита Ω. Дополнительно предполагаем, что в этом множестве заведомо содержится фиксированный нами читаемые два текста. Поэтому число выборок будет равным h^V-1. Какова вероятность $P_V(m)$ того, что будет выбрано m читаемых текстов? Случайная величина m распределена по гипергеометрическому распределению

36.

$$P_V(m) = \frac{\binom{|M(V)|^2 - 1}{m}\binom{|\Omega|^V - |M(V)|^2}{h^V - 1 - m}}{\binom{|\Omega|^V - 1}{h^C - 1}}$$

$$P_V(0) = \frac{\binom{|M(d)|^V - 2^{H\,dV}}{h^V}}{\binom{|M(d)|^V}{h^V}}$$

37. *Для модели 2* языка наша задача состоит в оценки мощности множества всех пар СТ длины D полученных с помощью *плотных парных* конкатенаций пар соседних d-грамм в последовательности множеств

$$\left\{\begin{matrix}\mathfrak{I}_{j_1}^{t(j_1)}\\ \mathfrak{I}_{j_2}^{t(j_2)}\end{matrix}\right\}\left\{\begin{matrix}\mathfrak{I}_{j_1+1}^{t(j_1+1)}\\ \mathfrak{I}_{j_2+1}^{t(j_2+1)}\end{matrix}\right\}\dots\left\{\begin{matrix}\mathfrak{I}_{j_1+D-1}^{t(j_1+D-1)}\\ \mathfrak{I}_{j_2+(V-1)d}^{t(D-1)}\end{matrix}\right\}, \qquad (**)$$

Здесь d-граммы содержательного текста $\mathfrak{I}=i_1i_2\dots i_{Vd+d-1}$ используются с зацеплением $i_ji_{j+1}\dots i_{j+d-1}, i_{j+1}\dots i_{j+d}$. Соседние столбцы $\left\{\begin{matrix}\mathfrak{I}_{j_1}^{t(j_1)}\\ \mathfrak{I}_{j_2}^{t(j_2)}\end{matrix}\right\}\left\{\begin{matrix}\mathfrak{I}_{j_1+1}^{t(j_1+1)}\\ \mathfrak{I}_{j_2+1}^{t(j_2+1)}\end{matrix}\right\}$ соответствуют выборкам пар d-грамм из пар содержательных текстов, соответствующим позициям $\begin{matrix}j_1\\ j_2\end{matrix}$ и $\begin{matrix}j_1+1\\ j_2+1\end{matrix}$. Считаем, что помощью *плотных парных* конкатенаций строятся пары содержательных текстов. При этом предполагаем, что **одно решение – пара СТ заведомо существует.**

38.Будем действовать аналогично использованным приемам для модели 1. Представим число содержательных текстов $|M_1(D)|$ длины D в алфавите I в виде $|M_1(D)|=|I|^{H_D \cdot D}$, Тогда число СТ в алфавите Ω равно $|I|^{2H_D \cdot D}$, Число всех последовательностей букв алфавита Ω длины D-d+1 равно $|\Omega|^{D-d+1}$. Глубина колонок равна h, число всех возможных последовательностей d-грамм алфавита I, построенных зигзагообразным чтением из колонок есть h^{D-d+1}. Мы определили новые параметры для решаемой задачи в модели 2 языка, чтобы использовать гипергеометрическое распределение вероятностей. А будет ли отражать реальность эта вероятностная модель для случайной величины ξ - множества всех пар СТ длины D полученных с помощью *плотных парных* конкатенаций пар соседних d-грамм в последовательности множеств (**)?

ПРОШУ ЗАДАВАТЬ ВОПРОСЫ