3 Задачи криптоанализа

3.1 Атаки

Рассмотрим криптосистему $C = \langle \mathcal{K}, \mathcal{X}, \mathcal{Y}, E, D \rangle$, которую использует Алиса и Боб и пытается скомпрометировать Виктор. В криптоанализе считается, что Виктор точно знает устройство C, ему неизвестен только ключ K, который используют Алиса и Боб. Данное предположение есть известный в криптологии *принцип Керкгоффса* — надежность криптосистемы определяется лишь секретностью ключа.

Виктор наблюдает за C, получая полную или частичную информацию о парах «открытый текст X_t — шифртекст Y_t»:

$$Y_t = E_K(X_t), \quad t = 1, \ldots, T, \quad X_t \in \mathcal{X}, \quad Y_t \in \mathcal{Y}.$$

По наблюдениям Виктору требуется решить следующую задачу:

(C1) определить ключ K преобразования E_K.

Зная K, Виктор может построить преобразование D_K и определить открытый текст $X = D_K(Y)$ по любому перехваченному $Y = E_K(X)$. Вообще говоря, Виктору не обязательно находить ключ, он может просто построить D_K:

(C2) не определяя K, найти D_K, т. е. построить алгоритм нахождения X_{T+1} по заданному $Y_{T+1} = E_K(X_{T+1})$.

В некоторых случаях Виктору даже не требуется выполнять расшифрование. Его задача — проверить, действительно ли Алиса и Боб используют криптосистему C (поддержка принципа Керкгоффса):

(C3) по парам «открытый текст X_t — шифртекст $Y_t = e(X_t)$» определить, является преобразование e преобразованием зашифрования криптосистемы C или нет: $e \stackrel{?}{\in} E = \{E_K \colon K \in \mathcal{K}\}$.

Задача C1 является самой сложной (и практически значимой), задача C3 самой простой. Как правило, если существуют способ решения одной задачи, то возникают подходы к решению и всех остальных. Методы решения задач криптоанализа называются *атаками*.

Выделяют следующие типы атак:

1) известны $\{Y_t\}$ и некоторые свойства открытого текста $\{X_t\}$ (*атака при известном шифртексте*);

2) известны $\{X_t\}$ и $\{Y_t\}$ (*атака при известном открытом тексте*);

3) криптоаналитику предоставлена возможность заранее выбрать значения $\{X_t\}$ (*атака при выбранном открытом тексте*);

4) криптоаналитик может задать значение X_t для каждого $t = 2, \ldots, T$, зная $Y_1, \ldots, Y_{t-1}$ (*атака при выбираемом открытом тексте*).

Пример 3.1 (формат). Сообщение X может иметь определенный формат, известный Виктору. Например, пакеты протокола HTTP, отправляемые сервером могут иметь фиксированные заголовки, за которыми следуют собственно данные:

```
HTTP/1.1 200 OK
Server: Apache/2.2.10 (Unix) PHP/5.2.6
Content-Type: text/html; charset=UTF-8
Content-Length: 67
Connection: close

<html><head><title>Hello,</title></head>
<body>Alice</body></html>
```

Пример 3.2 (Энигма). Во время второй мировой войны британская специальная служба, размещенная в Блетчипарке, проводила криптоанализ немецкой шифровальной машины Энигма. Примерные атаки при известном и выбранном открытом тексте имели вид:

1. Один из немецких шифровальщиков в начале сеанса связи использовал один и тот же тестовый открытый текст, известный британцам.
2. По агентурным каналам в немецкие подразделения доводилась информация (ложная или правдивая) о наличии мин в тех или иных районах. Последующие сообщения немцев обязательно содержали слово “minen” (мины, нем.). □

Пример 3.3 (GSM). В сетях связи GSM второго поколения речевые данные оцифровываются. Каждым 18.4 мс разговора соответствует двоичное слово длины 184. Для противодействия помехам в канале связи слово (как вектор-строка) умножается на двоичную матрицу размера 184×456. В результате получается кодовое слово X, которое обладает структурными особенностями: имеется $456 - 184$ независимых линейных комбинаций символов X, которые обязательно обращаются в $0 \bmod 2$. Кодовое слово X разбивается на 4 фрейма — слова длины 114. Каждый фрейм зашифровывается перед отправкой в канал связи. Виктор, который перехватывает зашифрованные фреймы, знает о стуктурных особенностях соответствующего открытого текста X (хотя не располагает информацией о самих речевых данных). □

Пример 3.4 (терминалы и карты). Продемонстрируем актуальность перечисленных атак на следующем примере. Пусть имеется система терминалов и карт. В защищенной памяти карты и терминала хранится секретный ключ K. Для того, что подтвердить подлинность карты, выполняются следующие действия:

1) терминал генерирует случайный открытый текст $X_t \in \mathcal{X}$, и отправляет его карте;

2) карта выполняет зашифрование и отправляет результат $Y_t = E_K(X_t)$ ттерминалу;

3) терминал проверяет совпадение $Y_t \stackrel{?}{=} E_K(X_t)$. Если проверка прошла успешно, то карта признается подлинной.

Как видим, располагая картой и необходимыми техническими средствами, злоумышленник может эмулировать терминал, отправлять карте произвольные тексты X_t и получать результаты их зашифрования Y_t. □

Сложность методов криптоанализа характеризуется следующими величинами:

1) количеством T пар (X_t, Y_t) (объем материала);

2) числом элементарных операций (время атаки);

3) объем необходимой памяти (память атаки).

При этом получаемые в ходе атаки решения зачастую носят вероятностный характер — мы определяем истинные значения битов ключа не наверняка, а с некоторой вероятностью p.

3.2 Частотные атаки

Далее мы подробно рассмотрим атаки при известном шифртексте. Пусть A — естественный алфавит, $L \subseteq A^*$ — естественный язык, Алиса и Боб обмениваются словами L, используя C.

Естественные языки обладают статистическими закономерностями, которые может использовать Виктор. Самые простые закономерности — разные частоты встречаемости символов, пар последовательных символов (биграмм), триграмм и т.д.

Например, 10 самых частых символов русского языка:

символ	частота
О	0.090
Е	0.072
А	0.062
И	0.062
Т	0.053
Н	0.053
С	0.045
Р	0.040
В	0.038
Л	0.035

10 самых частых биграмм: СТ, НО, ЕН, ТО, НА, ОВ, НИ, РА, ВО, КО.

Статистические свойства языка можно использовать для криптоанализа. Рассмотрим два примера.

Частоты символов. Пусть Алиса и Боб общаются по-русски, C — шифр сдвига. Виктор перехватил шифртекст

ВЩЦГЪБГЫВГЩЦГЪБГЫВГ.

Чаще всего в шифртексте встречается символ Г. С другой стороны, О — самый частый символ открытого текста. Виктор строит оценку ключа $\hat{K} = \text{Г}(3) - \text{О}(15) = \text{Ф}(21)$. Оценка оправдывается: при расшифровании на построенном ключе получается открытый текст

НЕВОЗМОЖНОЕВОЗМОЖНО.

Анализ частот встречаемости символов можно использовать для проведения атаки при известном открытом тексте на произвольный шифр простой замены. Для этого Виктор сортирует символы A двумя способами:

1) по убыванию их встречаемости в естественном языке и получает последовательность $a_1, a_2, \ldots, a_{|A|}$;
2) по убыванию их встречаемости в перехваченном шифртексте и получает последовательность $b_1, b_2, \ldots, b_{|A|}$.

Искомая таблица замены:

$$\hat{K} = \begin{pmatrix} a_1 & a_2 & \ldots & a_{|A|} \\ b_1 & b_2 & \ldots & b_{|A|} \end{pmatrix}.$$

Возможно, некоторые замены определены неверно. Виктор исправляет ошибки, дополнительно анализируя результат расшифрования.

Частоты биграмм. Пусть Алиса и Боб общаются по-русски, C — шифр перестановки с ключом $K \in S(\{1,2,3,4\})$. Пусть $K^{-1} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ a & b & c & d \end{pmatrix}$, т.е. при зашифровании выполняется преобразование

$$x_1x_2x_3x_4 \mapsto x_ax_bx_cx_d.$$

Виктор перехватил шифртекст

РАОП ДЬОЛ УТПС АНША ТСОЕ ТСРЙ ИНЦА ШВЕЕ ОТЙС РТОС ЕКТЧ СКТЕ.

Виктор разбивает шифртекст на фрагменты длины 4 и анализирует расположение символов С и Т. Имеются шаблоны: xTxC, TCxx, CxTx.

При зашифровании высоковероятные шаблоны открытого текста переходят во встреченные шаблоны шифтекста:

$x_1x_2x_3x_4$		$x_ax_bx_cx_d$
CTxx		xTxC
xCTx	$\mapsto$	TCxx
xxCT		CxTx

Анализ шаблонов показывает, что с высокой долей достоверности $b = d + 1$, $a = b + 1$, $c = a + 1$. Следовательно, $abcd = 3241$ и $K = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 2 & 1 & 3 \end{pmatrix}$. Действительно, после расшифрования на данном ключе получаем открытый текст

ПАРОЛЬДОСТУПАНАШЕСТОЙСТРАНИЦЕВШЕСТОЙСТРОЧКЕТЕКСТ.

Пример 3.5. Во время II мировой войны большинство американских шифровальщиков были индейцами племени навахо, которые обменивались словами своего родного редкого (!) языка. Свойства данного языка, естественно, не были известны противникам американцев — японцам. □

3.3 Криптоанализ шифра Виженера

Определение 3.1. Пусть $a = a_1 \dots a_n$ — слово в алфавите A. *Индексом совпадений* называется вероятность $I_c(a)$ совпадения двух наудачу выбранных символов a:

$$I_c(a) = \mathsf{P}\left\{a_i = a_j \colon i, j \overset{R}{\leftarrow} \{1, 2, \dots, n\}\right\}.$$

Свойства индекса совпадения:

1. Если $\nu_\alpha(a)$ — число символов α в слове a, то

$$\begin{aligned} I_c(a) &= [\mathbf{I}\{\mathcal{E}\} \text{ — индикатор наступления события } \mathcal{E}] = \\ &= \frac{1}{n^2} \sum_{1 \leqslant i,j \leqslant n} \mathbf{I}\{a_i = a_j\} = \\ &= \frac{1}{n^2} \sum_{1 \leqslant i,j \leqslant n} \sum_{\alpha \in A} \mathbf{I}\{a_i = \alpha\} \mathbf{I}\{a_j = \alpha\} = \\ &= \frac{1}{n^2} \sum_{\alpha \in A} \left(\sum_{i=1}^{n} \mathbf{I}\{a_i = \alpha\} \right) \left(\sum_{j=1}^{n} \mathbf{I}\{a_j = \alpha\} \right) = \\ &= \sum_{\alpha \in A} \left(\frac{\nu_\alpha(a)}{n} \right)^2. \end{aligned}$$

2. Пусть $L \subseteq A^*$ — язык, $a \overset{R}{\leftarrow} L$, p_α — вероятность появления символа α в слове a. Тогда

$$I_c(a) = \sum_{\alpha \in A} \left(\frac{\nu_\alpha(a)}{n} \right)^2 \approx \left[p_\alpha \approx \frac{\nu_\alpha(a)}{n} \right] \approx \sum_{\alpha \in A} p_\alpha^2.$$

3. Пусть $s \in S(A)$ — подстановка на A. Тогда

$$I_c(a_1 a_2 \dots a_n) = \Big[a_i = a_j \Leftrightarrow s(a_i) = s(a_j) \Big] = I_c(s(a_1) s(a_2) \dots s(a_n)).$$

4. Пусть $s_1, \dots, s_n \overset{R}{\leftarrow} S(A)$. Тогда

$$I_c(s_1(a_1) s_2(a_2) \dots s_n(a_n)) \approx \left[\frac{\nu_\alpha(s_1(a_1) s_2(a_2) \dots s_n(a_n))}{n} \approx \frac{1}{|A|} \right] \approx \sum_{\alpha \in A} \left(\frac{1}{|A|} \right)^2 = \frac{1}{|A|}.$$

Вернемся к шифру Виженера. Пусть $A = \mathbb{Z}_m$, $K \in A^*$ — ключ длины $l_0 > 0$, $X = x_1 \dots x_T \in L$ — открытый текст, $Y = y_1 \dots y_T$ — шифртекст.

Разобъем X и Y на фрагменты:

$$X_i(l) = x_i x_{i+l} \dots x_{i+ld}, \quad Y_i(l) = y_i y_{i+l} \dots y_{i+ld}, \quad i = 1, \dots, l,$$

где $d = d(i)$ — максимальное целое, такое что $i + ld \leqslant T$.

Рассмотрим два случая.

А. $l = l_0$. В этом случае $Y_i(l)$ получается из $X_i(l)$ с помощью шифра сдвига на ключе K_i, $i = 1, 2, \dots, l$. Индекс совпадений

$$I_c(Y_i(l)) \overset{3}{=} I_c(X_i(l)) \overset{2}{\approx} \sum_{\alpha \in A} p_\alpha^2.$$

Например, для русского языка индекс совпадений: ≈ 0.0529.

B. $l \neq l_0$. В этом случае символы $Y_i(l)$ получаются из символов $X_i(l)$ с помощью сдвигов на различные символы ключа K. При этом символы $Y_i(l)$ становятся более «случайными», чем в предыдущем случае. В идеале: символы $Y_i(l)$ являются результатом применения к символам $X_i(l)$ различных случайных подстановок. Индекс совпадений

$$I_c(Y_i(l)) \overset{4}{\approx} \frac{1}{m}.$$

Например, для русского языка: ≈ 0.0303.

Разница между индексами совпадений позволяет Виктору провести следующую атаку:

1. Просмотреть $l = 1, 2, \ldots, l_{max}$.
2. Для каждого l составить фрагменты $Y_i(l)$ и найти индексы совпадений.
3. Выбрать l, для которого индексы совпадений максимальны.
4. Определить ключи K_i шифров сдвига $X_i(l) \mapsto Y_i(l)$, используя частотные свойства языка.

Отметим, что Виктор применяет важный криптоаналитический принцип: divide et impera (разделяй и властвуй). Виктор определяет ключ K по частям: сначала его длину, а затем каждый символ по отдельности.