НАДЕЖНЫЕ СИСТЕМЫ ЗАЩИТЫ ЭЛЕКТРОННЫХ ПУБЛИКАЦИЙ, БАЗИРУЮЩИЕСЯ НА ЭФФЕКТИВНОМ ОМОФОННОМ КОДИРОВАНИИ*

Б. Я. Рябко, А. Н. Фионов
Сибирская государственная академия телекоммуникаций и информатики
Новосибирск, Россия
e-mail: ryabko@adm.ict.nsc.ru

А. М. Федотов
Институт вычислительных технологий СО РАН
Новосибирск, Россия
e-mail: fedotov@adm.ict.nsc.ru

The paper deals with the problem of the protection (security) of electronic publications from unauthorized access (reading), e. g., in the course of data transfer through open communication lines, organization of authorized access and access to the Internet data bases.

Проблема защиты электронных публикаций имеет множество аспектов. Здесь необходимо решать такие задачи, как предотвращение несанкционированного доступа, контроль целостности и аутентичности информации, подтверждение авторства и обеспечение авторских прав и др. Легкость копирования, перезаписи и дальнейшего распространения авторских материалов, опубликованных в электронном виде, увеличивает количество злоупотреблений по сравнению с обычными печатными изданиями. Проблему защиты электронных изданий можно разделить на две: первая связана с системой защиты информации на сервере, вторая — с перехватом информации при ее передаче по открытым сетям. Такое деление в основном связано с методами, используемыми для защиты данных. Если первая проблема может быть решена путем организации авторизованного доступа к данным, то решить вторую можно только с привлечением методов кодирования и шифрования информации.

В настоящей работе мы остановимся на проблеме защиты текстов электронных публикаций от несанкционированного прочтения посторонними лицами, например, при передаче данных по открытым каналам связи. Эта задача может быть решена путем шифрования данных и целью работы является представление класса методов, позволяющих строить эффективные и надежные шифры, базирующиеся на криптографических алгоритмах.

Важнейшим параметром любой криптосистемы является ее стойкость, т.е. способность противостоять различного рода атакам, сохраняя секретность. Мы рассматриваем только криптологические атаки, когда криптоаналитику, пытающемуся вскрыть шифр, известны шифротекст, метод шифрования и статистические особенности зашифрованного сообщения, но не известен секретный ключ.

Практически все используемые в настоящее время криптосистемы (DES, IDEA, RSA, DSS, ГОСТ 28147-89 и др.) являются лишь вычислительно стойкими. Это означает, что их стойкость основывается на сложности алгоритма, позволяющего раскрыть зашифрованное сообщение только на основе шифротекста (без знания секретного ключа). Такой алгоритм существует, его можно реализовать, например, путем простого перебора ключей. С теоретико-информационной точки зрения это означает, что шифротекст содержит в себе информацию о зашифрованном сообщении. Показано, в частности, что для английского текста, зашифрованного алгоритмом DES с длиной ключа 56 бит, достаточно в среднем 11 байт шифротекста, чтобы однозначно восстановить зашифрованное сообщение. Конечно, переборный алгоритм не может быть использован на практике (он имеет экспоненциальную сложность, т.к. количество просматриваемых вариантов ключей экспоненциально возрастает при увеличении размера ключа). Основным вопросом, широко исследуемым в настоящее время, является следующий: имеется ли для конкретной вычислительно стойкой криптосистемы алгоритм взлома, существенно более эффективный, чем перебор ключей. К сожалению, ответ на этот вопрос не известен. Правда, не известны (по крайней мере, широкой научной общественности) и эффективные алгоритмы взлома вышеперечисленных криптосистем. Основным достижением теории сложности алгоритмов является выделение класса так называемых НП-полных задач. К этому классу, в частности, относятся многие известные в течение столетий проблемы, для которых до сих пор не найдены эффективные методы решения. Несмотря на то, что принципиальная труднорешаемость НП-полных задач также не доказана, установление принадлежности алгоритма вскрытия шифра классу НП давало бы некоторую гарантию стойкости криптосистемы. Для большинства современных систем, однако, принадлежность к классу НП либо не доказана, либо, наоборот, доказано, что они не принадлежат к этому классу (например, метод RSA). После того как для некоторых НП-полных задач, например ранцевой криптосистемы, были построены эффективные вероятностные алгоритмы, дающие быстрый ответ если не во всех, то в большинстве случаев, ситуация с труднорешаемостью еще более осложнилась. Таким образом, на сегодняшний день положение дел таково, что в надежность вычислительно стойких криптосистем можно только верить.

На другом полюсе по отношению к вычислительно стойким системам стоят безусловно стойкие криптосистемы. Стойкость этих систем не зависит от того, какими вычислительными возможностями обладает криптоаналитик. Для безусловно стойких систем просто не существует алгоритма взлома (в связи с чем отпадает вопрос о его сложности). Одной из самых старых безусловно стойких криптосистем является шифр Вернама: символы сообщения складываются с символами ключа, причем ключ является одноразовым (т.е. используется только один раз) и его длина равна длине сообщения. Для этого шифра не существует алгоритма взлома. При переборе ключей мы будем получать все возможные варианты сообщений источника и нет никакой дополнительной информации, которая позволила бы отдать предпочтение тому или иному варианту. Применение такого шифра на практике весьма затруднительно, т. к. требуется наличие длинных секретных ключей у отправителя и получателя сообщений. На языке теории информации это означает, что

пропускная способность закрытого канала должна быть равна пропускной способности открытого канала, что можно позволить только в самых исключительных случаях (например, шифр Вернама используется в межправительственной связи Москва — Вашингтон).

Другим подходом к построению безусловно стойких систем является сжатие данных и рандомизация. Сжатие данных означает устранение избыточности. Если избыточность равна нулю, то любая кодовая комбинация соответствует какому-либо сообщению. Тогда даже при коротком ключе количество вариантов расшифровки равно количеству всевозможных ключей и шифр становится невскрываемым. На практике идеальное сжатие недостижимо хотя бы по причине конечности кодирующего алфавита, а также из-за ограничений, вызываемых вычислительной сложностью алгоритмов кодирования и декодирования. Реально можно говорить не о нулевой, но заданной произвольно низкой избыточности. Ненулевая избыточность приводит к “старению"ключа. По мере кодирования символов источника все больше информации о ключе открывается в шифротексте, количество вариантов расшифровки снижается и ключ необходимо обновлять. Темп обновления ключа оказывается существенно меньшим, чем в шифре Вернама (и зависит от величины избыточности). Заметим, что использование стандартных архиваторов в криптографических целях не имеет смысла, так как все они являются адаптивными кодерами и фактически начинают сжимать данные только после нескольких сотен символов сообщения. Но, как уже было сказано, бывает достаточно нескольких десятков байт шифротекста, чтобы восстановить использованный ключ. Поэтому универсальные кодеры должны быть интегрированы с шифратором в одной системе для того, чтобы интенсивность использования ключевой информации была согласована со степенью сжатия.

При рандомизации к кодовой последовательности определенным образом “примешивается"шум от дополнительного шумового источника (т. е. источника случайных чисел). В результате, даже при высокой избыточности кода, кодовую последовательность удается сделать полностью случайной, т. е. неотличимой от последовательности равновероятных и независимых кодовых символов (ниже будут приведены примеры рандомизации на основе так называемого омофонного кодирования). Несмотря на то, что количество вариантов расшифровки при рандомизации может быть существенно меньше мощности множества ключей, информация о самом ключе остается полностью закрытой (просто отсутствует в шифротексте) независимо от длины сообщения. Это означает, что ключ не устаревает по мере кодирования.

Идеальным решением является объединение сжатия и рандомизации. В этом случае мы получаем заданную высокую стойкость шифра при коротком неустаревающем секретном ключе. Нами разработан класс методов сжимающей рандомизации, обеспечивающих полную случайность кодовой последовательности при заданных произвольно низких избыточности и количестве используемых случайных символов. По своей вычислительной эффективности эти методы существенно превосходят все известные методы, требуя лишь линейного или даже логарифмического роста объема памяти кодера и декодера и полиномиально-логарифмического роста времени кодирования и декодирования при снижении избыточности.

Поясним суть методов рандомизации на следующих примерах. Пусть бернуллиевский источник порождает буквы в алфавите $A = \{a, b\}$ с вероятностями $P(a) = 3/4$, $P(b) = 1/4$. Будем кодировать сообщение $aaba$.

Традиционная схема омофонного кодирования описывается следующим образом:

$$a \to \begin{cases} 00, & \text{с вероятностью } 1/3, \\ 01, & \text{с вероятностью } 1/3, \\ 10, & \text{с вероятностью } 1/3, \end{cases}$$

$$b \to \{11, \text{ с вероятностью } 1.$$

Так как буква a может кодироваться тремя кодовыми словами, выбираемыми с одинаковой вероятностью, мы можем получать различные варианты кодов сообщения. Все они будут иметь длину 8 бит. Запишем один из них:

$$C(aaba) = 00011110.$$

Развитием этого подхода является схема Гюнтера:

$$a \to \begin{cases} 0, & \text{с вероятностью } 2/3, \\ 10, & \text{с вероятностью } 1/3, \end{cases}$$

$$b \to \{11, \text{ с вероятностью } 1.$$

Средняя длина кода сообщения составляет 6 бит, и одно из возможных кодовых слов

$$C(aaba) = 010110.$$

Нами предлагаются следующие два базовых метода.

Метод блокового омофонного кодирования. Определим кумулятивные вероятности

$$Q(a) = 0, Q(b) = Q(a) + P(a) = 3/4.$$

Для определения кодовых слов необходимо вычислить интервал $[Q, \tilde{Q}]$, соответствующий сообщению $U = aaba$ на единичном отрезке $[0, 1)$. Для этого необходимо вычислить

$$q_1^1 = Q(u_1) + P(u_1)Q(u_2) = 0 + \frac{3}{4} \cdot 0 = 0,$$

$$p_1^1 = P(u_1)P(u_2) = \frac{3}{4} \cdot \frac{3}{4} = \frac{9}{16},$$

$$q_2^1 = Q(u_3) + P(u_3)Q(u_4) = \frac{3}{4} + \frac{1}{4} \cdot 0 = \frac{3}{4},$$

$$p_2^1 = P(u_3)P(u_4) = \frac{1}{4} \cdot \frac{3}{4} = \frac{3}{16},$$

$$q_1^2 = q_1^1 + p_1^1 q_2^1 = 0 + \frac{9}{16} \cdot \frac{3}{4} = \frac{27}{64},$$

$$p_1^2 = p_1^1 p_2^1 = \frac{9}{16} \cdot \frac{3}{16} = \frac{27}{256}.$$

Тогда границы интервала

$$Q = q_1^2 = \frac{108}{256} = 0,01101100,$$

$$\tilde{Q} = q_1^2 + p_1^2 - 1 = \frac{134}{256} = 0,1000011.$$

Теперь достаточно выбрать случайным образом одно из кодовых слов в интервале $[Q, \tilde{Q}]$. Наиболее вероятным будет кодовое слово

$$C(aaba) = 0111.$$

Метод арифметического кодирования с разделением интервала. В этом методе также вычисляется интервал для всего сообщения, но на каждом шаге может проводиться разделение интервала и масштабирование. Схематично процесс кодирования может быть представлен следующим образом:

$$[0,1) \xrightarrow{a} \left[0, \frac{3}{4}\right) \xrightarrow{a} \left[0, \frac{9}{16}\right) \to \begin{cases} \left[0, \frac{8}{16}\right) & \text{с вероятностью } 8/9, \\ \left[\frac{8}{16}, \frac{9}{16}\right) & \text{с вероятностью } 1/9. \end{cases}$$

Допустим, выбран более вероятный интервал $\left[0, \frac{8}{16}\right) = \left[0, \frac{1}{2}\right)$. Продолжим процесс:

$$\left[0, \frac{1}{2}\right) \xrightarrow[0]{} [0,1) \xrightarrow{b} \left[\frac{3}{4}, 1\right) \xrightarrow[11]{} [0,1) \xrightarrow{a} \left[0, \frac{3}{4}\right) \to \begin{cases} 0 \text{ с вероятностью } 2/3, \\ 10 \text{ с вероятностью } 1/3. \end{cases}$$

Здесь над стрелкой показан символ сообщения, в соответствии с которым сужается интервал, а под стрелкой показаны кодовые биты, формируемые в результате масштабирования. Построение кодового слова завершается омофонным кодированием заключительного интервала. Наиболее вероятное кодовое слово

$$C(aaba) = 0110.$$

При использовании рандомизации метод шифрования может быть очень простым (например, посимвольное сложение закодированной последовательности с секретным ключом, как в шифре Вернама). Поэтому по скорости работы предлагаемые нами методы не уступают вычислительно стойким системам. Так, один из наиболее быстрых блочных шифров IDEA требует 8 операций умножения на каждый байт сообщения, тогда как метод арифметического кодирования с разделением интервала требует только двух умножений и делений на каждый байт сообщения. При этом методы сжимающей рандомизации обеспечивают безусловную стойкость криптосистемы при использовании коротких многоразовых секретных ключей.

Поступила в редакцию 24 апреля 1997 г.