АТАКА НА ПОТОКОВЫЕ ШИФРЫ RC4 И ZK-CRYPT С ИСПОЛЬЗОВАНИЕМ ТЕСТА «СТОПКА КНИГ»

С. А. Дорошенко, А. М. Лубкин, В. А. Монарев, Б. Я. Рябко, А. Н. Фионов

Недавно предложенный статистический тест, названный «Стопка книг», был успешно применён для нахождения отклонений от случайности в выходной последовательности бит потоковых шифров RC4 и ZK-Crypt. Приведено краткое описание теста и результаты экспериментов.

1. Введение

Цель так называемой атаки различения (distinguishing attack) на потоковые шифры и генераторы псевдослучайных бит – найти отклонения от случайности в выходной последовательности. Принято считать, что генератор (и соответствующий потоковый шифр, основанный на нем) криптографически стоек, если он проходит все статистические тесты, т.е. битовая последовательность, получаемая на выходе генератора, неотличима от истинно случайной последовательности с помощью этих тестов (см. [1]). С другой стороны, если тест нашел (за полиномиальное время) отклонения от случайности, то такой генератор (шифр) не рекомендуют к применению в криптографии. Поэтому атака такого типа очень важна.

Существуют два основных подхода к реализации атаки различения. Первый подход основан на нахождении отклонений от случайности с помощью теоретического исследования генератора. Второй подход использует результаты статистических тестов. Этот подход широко известен и часто применяется на практике. В частности, Национальный институт стандартов и технологий США (NIST) рекомендовал в [2] этот подход и предложил 16 статистических тестов для применения к криптографическим генераторам. В [2] можно также найти документацию к предложенным статистическим тестам. Мы кратко изложим основную идею нашего теста.

Обозначим через H_0 гипотезу о равномерном распределении бит выходной последовательности, а альтернативную ей гипотезу – через H_1. С помощью статистического теста принимается та или иная гипотеза. Более точно, H_0– это гипотеза о том, что последовательность бит порождена источником, выход которого независим и одинаково распределён (в двоичном случае источник порождает нули и единицы независимо и с вероятностями $^1/_2$). Альтернативная гипотеза $H_1 = \neg H_0$ состоит в том, что последовательность сгенерирована стационарным эргодическим источником, отличным от источника при выполнении H_0. В статистике тестируемую последовательность называют выборкой. Часто всю последовательность делят на отдельные независимые последовательности для тестирования.

Проверка гипотез основана на методах математической статистики. Напомним, что ошибкой первого рода называют вероятность отвергнуть H_0 при её выполнении, обозначим её через α. Величину $1-\alpha$ называют уровнем значимости. Обычно значение величины α лежит в диапазоне от 0.001 до 0.05. В аналогичной ситуации, если верна H_1, то вероятность отвергнуть гипотезу H_1 называют ошибкой второго рода и обозначают через β. Как правило, величину β сложно определить. Заметим, что последовательность, которая сгенерирована некоторым генератором псевдослучайных бит, не может быть случайной, поскольку генератор – это некоторый детерминированный алгоритм. С другой стороны, для хороших (криптостойких) генераторов любой статистический тест должен принимать гипотезу H_0.

В [3] был предложен статистический тест «Стопка книг». Он был описан в терминах математической статистики (см. также [4]). В ряде работ авторами было показано, что в некоторых случаях этот тест оказывается более мощным, чем все 16 тестов, предложенных NIST. В этой статье мы представляем результаты применения «Стопки книг» для тестирования генераторов RC4 и Zk-Crypt.

RC4 – это хорошо известный криптографический генератор. В работах [5, 6, 7, 8] было показано, что последовательность, которую генерирует RC4, неслучайна. В [9] неслучайность RC4 была экспериментально исследована на выборке более чем 2^{55} бит. Мы рассмотрим наиболее распространённый вариант RC4, в котором число бит на один выход шифра равно восьми. В наших экспериментах показано, что тест «Стопка книг» находит отклонения от случайности у RC4 на выборке всего 2^{32} бит.

Zk-Crypt [10] – это новый потоковый генератор, один из кандидатов конкурса ECRYPT stream cipher project (eSTREAM). Нами показано, что «Стопка книг» находит отклонения от случайности на выборке 2^{55} бит.

Можно заметить, что для атаки с помощью «Стопки книги» на шифры RC4 и Zk-Crypt достаточно обычного персонального компьютера.

Важно также отметить, что остальные кандидаты eStream прошли тест на выборке 2^{30}–2^{35} бит (объём выборки зависит от скорости шифра и наших вычислительных возможностей).

Последнее замечание касается названия теста. В универсальном кодировании известен метод «Стопка книг», который был предложен Б.Я. Рябко [11]. Через несколько лет он был переоткрыт в [12] и назван «Move-To-Front» (MTF). Последнее название было принято множеством западных исследователей, незнакомых с опубликованным ранее русскоязычным источником.

Компьютерная реализация стопки «Стопки книг» доступна на сайте http://web.ict.nsc.ru/~ldp.

Работа организована следующим образом. В разделе 2 приводятся некоторые факты из математической статистики. В разделе 3 приводится описание теста «Стопка книг». В третьем разделе приведены результаты экспериментов.

2. Статистический критерий

В нашем тесте мы используем широко известный критерий хи-квадрат, который может быть описан следующим образом. Пусть n – объем выборки, n_0 – число нулей, n_1 – число единиц в выборке, $n = n_0 + n_1$. Обозначим через p и q априорные вероятности нуля и единицы соответственно. Тогда np и nq – это ожидаемое число нулей и единиц в выборке размера n. Далее вычисляем статистику x^2:

$$x^2 = \frac{(n_0 - pn)^2}{pn} + \frac{(n_1 - qn)^2}{qn} \qquad (1)$$

Для проверки гипотезы H_0 имеем $p=q={}^1\!/_2$ и из (1) получаем

$$x^2 = \frac{(n_0 - n_1)^2}{pn}.$$

Однако прямое тестирование обычно не достаточно эффективно и, как правило, используются некоторые трансформации тестируемой последовательности с последующей проверкой иных, но эквивалентных, гипотез.

Схема статистического теста следующая. Мы задаём критическое значение $t_\alpha > 0$. Затем вычисляем по выборке статистику x^2 и сравниваем с критическим значением. Принимаем гипотезу если, $x^2 \ge t_\alpha$, иначе гипотеза отвергается. Таким образом, ошибка первого рода – это вероятность события $x^2 > t_\alpha$, когда верна H_0, то есть $\alpha = P(x^2 > t_\alpha)$.

Известно, что указанная статистика x^2 подчиняется распределению хи-квадрат с одной степенью свободы. Также известно, что тест хи-квадрат применим, если оба значения и больше пяти. В экспериментах мы использовали вероятности ошибки первого рода 0.05 и 0.001, критические значения для которых можно найти в литературе (см. [1]):

$$\alpha = 0.05 \qquad t_\alpha = 3.8415$$
$$\alpha = 0.001 \qquad t_\alpha = 10.8376$$

Например, если в серии экспериментов (на независимых выборках) при вероятности ошибки 0.05 значение статистики больше, чем 3.8415, мы делаем вывод, что последовательность неслучайна (H_0 отвергается).

3. Стопка книг

В тесте «стопка книг», который был предложен для атаки, входная последовательность $x_1, x_2, \ldots, x_S$ размера S, где $x_i \in \{0,1\}$, рассматривается как последовательность слов длинны w ($1 \le w \le 32$), образованных подряд идущими битами, между которыми есть промежутки (пропущенные биты или пробелы). Например, если $w = 3$ и пробел равен 1 бит, то последовательность 0111010100010100... преобразуется в последовательность слов 011 010 000 010 ... или, в десятичной записи, 3 2 0 2 ...

Таким образом, мы имеем выборку $y_1, y_2,..., y_n$, полученную из исходной последовательности, где $y_i \in \{0,1,\ldots,2^w - 1\}$.

В тесте «Стопка книг», все слова упорядочены. Обозначим порядковый номер слова a через $v(a)$. Порядок изменяется после обработки очередного символа по следующему правилу:

$$v^{i+1}(a) = \begin{cases} 0, & \text{если } y_i = a, \\ v^i(a) + 1, & \text{если } v^i(a) < v^i(y_i), \\ v^i(a), & \text{если } v^i(a) > v^i(y_i), \end{cases} \qquad (2)$$

где v^i – порядок после обработки $y_1, y_2, \ldots, y_i$, i=1, K n. Первоначальный порядок произвольный, например, $v^1 = \{0, 1, \mathrm{K}, 2^w - 1\}$.

Поясним (2) неформально. Полагаем, что слова помещены в стопку (по аналогии с книжной стопкой) и $v^1(a)$ – положение слова a в стопке. Пусть первое слово y_1 в $y_1, y_2, \ldots, y_n$ было a. Если это произошло, когда номер a был равен k ($v^1(a) = k$), то извлекаем a из стопки и перемещаем его наверх (это означает, что порядок изменился в соответствии с (2)). Повторяем эту же процедуру для y_2 и т. д.

Основную идею теста поможет понять тот факт, что при выполнении гипотезы H_1 наиболее часто встречаемые слова (как и книги в стопке) будут иметь наименьшие порядковые номера в стопке, если же выполнена гипотеза H_0, то вероятность, что произвольное слово займет произвольную позицию, равна $\frac{1}{2^w}$.

Продолжим описание теста. Множество всех номеров $\{0, 1, \ldots, 2^w - 1\}$ разбивается на две части $A_0 = \{0, 1, \ldots, u-1\}$ и $A_1 = \{u, \ldots, 2^w - 1\}$. Множество A_0 будем называть верхней частью стопки книг, его мощность равна $|A_0| = u$. За-

тем, обрабатывая y_1, y_2,..., y_n, мы подсчитываем, сколько $v^i(y^i)$, $i = 1, \ldots n$ принадлежит множествам A_0 и A_1. Обозначим эти числа через n_0 и n_1. Более формально:

$$n_k = |\{i : v^i(y_i) \in A_k, i = 1,\ldots,n\}|, \quad k = 0,1$$

Очевидно, что если выполнена гипотеза H_0, то все слова появляются с вероятностью $1/2^w$ и вероятность события $v^i(y_i) \in A_k$ равна $|A_k|/2^w$. Используя формулы раздела 2 и $|A_0| = u$, можно написать, что $p=u/2^w$ и $q=1-p$. Теперь проверка гипотезы H_0 заменена проверкой эквивалентной гипотезы H^*_0 о том, что бинарная последовательность случайных величин Y имеет распределение $P(Y = 0)$, $P(Y = 1) = q$ для выборки y_1, y_2,..., y_n (число нулей и единиц равно n_0 и n_1 соответственно). Для этого можно использовать тест хи-квадрат, который был описан во втором разделе.

Мы не даём описание подбора параметров тестирования, а именно, длину слова w, число пропущенных бит и размер верхней части u, но рекомендуем производить специальные эксперименты для подбора параметров для минимизации размера выборки. Существует множество криптографических приложений, в которых можно провести несколько экспериментов, оптимизирующих параметры, а затем проверять гипотезу на независимых от экспериментов данных. Например, для тестирования генератора возможен поиск параметров по одной части последовательности, а проверка гипотезы – по другой части.

Рассмотрим пример:

$w = 3$, $\quad y_1,..,y_8 = 3\ 6\ 3\ 3\ 6\ 1\ 6\ 1$

$u = 3$, $\quad v^1 = (0, 1, 2, 3, 4, 5, 6, 7).$

Тогда

$v^1 = (0, 1, 2, 3, 4, 5, 6, 7),\ n_0 = 0,\ n_1 = 0;.$

$v^2 = (3, 0, 1, 2, 4, 5, 6, 7),\ n_0 = 1;$

$v^3 = (6, 3, 0, 1, 2, 4, 5,\ 7),\ n_1 = 1;$

$v^4 = (3, 6, 0, 1, 2, 4, 5, 7),\ n_0 = 2;$

$v^5 = (3, 6, 0, 1, 2, 4, 5, 7),\ n_0 = 3;$

$v^6 = (6, 3, 0, 1, 2, 4, 5, 7),\ n_0 = 4;$

$v^7 = (1, 6, 3, 0, 2, 4, 5, 7),\ n_1 = 2;$

$v^8 = (6, 1, 3, 0, 2, 4, 5, 7),\ n_0 = 5;$

$v^9 = (1, 6, 3, 0\ , 2, 4, 5, 7),\ n_0 = 6.$

Можно заметить, что слова 3 и 6 встречаются достаточно часто, и это отклонение тест фиксирует достаточно эффективно. Действительно, среднее значение n_0 и n_1 равны 3 и 5, в то время как эксперимент показал значения 6 и 2 соответственно.

Остановимся на вопросе о сложности теста. «Наивная» реализация метода «Стопка книг» потребовала бы операций пропорционально 2^w для каждого элемента выборки. Если заметить, что достаточно хранить только верхнюю часть, то необходимое число операций составит $O(u)$. Более эффективные методы, основанные на AVL деревьях или других сбалансированных деревьях, требуют $O(\log u)$ операций. В нашей реализации используются ещё более быстрые алгоритмы – алгоритмы хеширования, для которых время выполнения составляет $O(1)$.

4. Реализация атаки

Случайность генератора RC4 была проверена для 100 случайных ключей с помощью теста «Стопка книг». Сначала вся битовая последовательность разбивается на слова длины 16 ($w = 16$, без пропусков) и верхняя часть берётся размера $u = 16$. Генерируем файлы различной длины для каждого случайного 128-битового ключа (см. Таблицу 1) и применяем «Стопку книг» для каждого файла с уровнем значимости 0.95. Таким образом, гипотеза H_0 выполняется, если в среднем 5 файлов из 100 будут признаны неслучайными. Все результаты приведены в таблице, где показано число файлов, которые были признаны неслучайными при данной длине выборки (из 100 файлов).

Таблица 1: тестирование RC4

Выборка (в битах)	2^{31}	2^{32}	2^{33}	2^{34}	2^{35}	2^{36}	2^{37}	2^{38}	2^{39}
Неслучайные	6	12	17	18	37	74	95	99	100

Учитывая, что всего 5 файлов из 100 могут быть признаны неслучайными, делаем вывод, что при длине выборки 2^{32} (и больше) последовательность можно отличить от случайной.

Случайность шифра Zk-Crypt также исследовалась для 100 случайных ключей. Выход шифра разбивался на слова длины 32 без пропусков, размер верхней части $u=2^{16}$. Для каждого ключа генерировали файл (см. Таблицу 2) и применяли «Стопку книг» с уровнем значимости 0.999. Если H_0 выполняется, то из 1000 файлов в среднем 1 файл будет признан неслучайным. Полученные результаты приведены в таблице. Видим, что генератор признан неслучайным на выборке примерно 2^{35}.

Таблица 2: тестирование Zk-Crypt

Выборка (в битах)	2^{24}	2^{26}	2^{28}	2^{30}
Неслучайные	25	51	97	100

Литература

1. Menezes A., van Oorschot P., Vanstone S. Handbook of Applied Cryptography. – CRC Press, 1996.

2. Rukhin A. et al. A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications. – NIST Special Publication 800-22 (rev. May 15, 2001).
3. Рябко Б.Я., Пестунов А.И. (2004). «Стопка книг» как новый статистический тест для случайных чисел // Проблемы передачи информации. – 2004. – т. 40, № 1. – с. 66-71.
4. Рябко Б.Я., Фионов А.Н. Методы криптографической защиты информации. – М.: Горячая линия – Телеком, 2006.
5. Dawson E., Gustafson H., Henricksen M., Millan B. (2002). Evaluation of RC4 stream cipher. – 2002. – http://www.ipa.go.jp/security/enc/CRYPTREC/fy15/.
6. Golic J. Dj. (2000). Iterative probabilistic cryptanalysis of RC4 keystream generator // Australasian Conference on Information Security and Privacy (ACISP) 2000. – p. 220-233.
7. Fluhrer S., McGrew D. Statistical analysis of the alleged RC4 keystream generator source // Proceedings of the 7th International Workshop on Fast Software Encryption. – 2000. (Lecture Notes in Computer Science; Vol. 1978). – p. 19-30.
8. Pudovkina M. Statistical weaknesses in the alleged RC4 keystream generator // Cryptology ePrint Archive. – 2002. – Report 2002/171. – (http://eprint.iacr.org/2002/171).
9. Crowley P. Small bias in RC4 experimentally verified. – (http://www.ciphergoth.org/crypto/rc4/).
10. Gressel C., Granot R., Vago G. ZK-Crypt // eSTREAM, ECRYPT Stream Cipher Project. – Report 2005. – (http://www.ecrypt.eu.org/stream/zkcrypt.html).
11. Рябко Б.Я. (1980). Сжатие информации с помощью «стопки книг» // Проблемы передачи информации. – 1980. – т. 16, № 4. – с. 16-21.
12. Bently J. L., Sleator D. D., Tarjan R. E., Wei V. K. A locally adaptive data compression scheme // Communications of the ACM. – 1986. – V. 29. – p. 320-330.

Дорошенко Сергей Александрович: аспирант кафедры прикладной математики и кибернетики, СибГУТИ, тел. (383) 266-71-35

Лубкин Алексей Михайлович: аспирант кафедры прикладной математики и кибернетики, СибГУТИ, тел. (383) 266-71-35

Монарев Виктор Александрович: к.ф.-м.н., доцент кафедры прикладной математики и кибернетики, СибГУТИ, тел. (383) 266-71-35

Рябко Борис Яковлевич: д.т.н., профессор, академик МАИ, проректор по научной работе СибГУТИ, зав. кафедрой прикладной математики и кибернетики, тел. (383) 266-82-23, e-mail: boris@ryabko.net

Фионов Андрей Николаевич: д.т.н., профессор кафедры прикладной математики и кибернетики, СибГУТИ, тел. (383) 266-71-35, e-mail: fionov@neic.nsk.su