ПРИБОРЫ И МЕТОДЫ ИЗМЕРЕНИЯ

УДК 519.2, 004.622, 004.056.53 DOI 10.21685/2307-5538-2019-3-4

В. И. Волчихин, А. И. Иванов, А. П. Карпов, А. П. Юнин

УСЛОВИЯ КОРРЕКТНОГО ВЫЧИСЛЕНИЯ ЭНТРОПИИ ОСМЫСЛЕННЫХ ДЛИННЫХ ПАРОЛЕЙ В ПРОСТРАНСТВЕ СВЕРТОК ХЭММИНГА С ЭТАЛОННЫМИ ТЕКСТАМИ НА РУССКОМ И АНГЛИЙСКОМ ЯЗЫКАХ

V. I. Volchikhin, A. I. Ivanov, A. P. Karpov, A. P. Yunin

CONDITIONS FOR THE CORRECT CALCULATION OF THE ENTROPY OF MEANINGFUL LONG PASSWORDS IN THE HAMMING CONVOLUTION SPACE WITH REFERENCE TEXTS IN RUSSIAN AND ENGLISH

А н н о т а ц и я. ***Актуальность и цели.*** Целью работы является повышение корректности вычисления энтропии длинных кодов с зависимыми разрядами, являющимися осмысленными легко запоминаемыми паролями на родном языке пользователя. ***Материалы и методы.*** Классические процедуры Шеннона не могут быть использованы, так как требуют использования огромного статистического материала. Для сокращения затрат вычислительных ресурсов используется отображение кодов в нормированное пространство сверток Хэмминга. ***Результаты.*** Показано, что результаты вычислений являются более корректными, если отказаться от побитного сложения по модулю 2 при вычислении сверток Хэмминга. Предложено использовать свертывание данных по модулю 8, так как кодирование паролей и эталонных текстов выполняется в 8-битной кодировке. Более того, корректное преобразование данных может быть выполнено только при использовании кода длинного пароля, свертываемого с эталонным текстом на родном языке пользователя. ***Выводы.*** В пространстве сверток Хэмминга легко вычислим прирост стойкости длинных легко запоминаемых паролей со смыслом к атакам подбора, возникающего из-за периодической смены регистра ввода длинного пароля.

A b s t r a c t. ***Background.*** The aim of the work is to increase the correctness of calculating the entropy of long codes with dependent bits, which are meaningful easily remembered passwords in the user's native language. ***Materials and methods.*** Since the classical Shannon procedures require the use of huge statistical material, they cannot be used in conditions of limited computing resources. To reduce the cost of computing resources, this paper uses code mapping in the normalized space of Hamming convolutions. ***Results***. The authors show that the results of computing the entropy of codes are more correct if one refuses bitwise addition modulo two when computing Hamming convolutions. In the article, it is suggested to use the data collapsing on module 8, since the encoding of passwords and reference texts are performed in 8-bit encoding. Moreover, the correct data conversion can be performed only by using a long password code that is collapsed with the reference text in the user's native language.

Conclusions. In the Hamming convolution space, it is easy to calculate the increase in the persistence of long, easy-to-remember passwords with meaning to brute-force attacks that arise from the periodic change of the long-password input register.

Ключевые слова: энтропия длинных кодов с зависимыми разрядами, регуляризация вычислений, многообразие сверток Хэмминга, требования к перекодировке данных перед их свертыванием по Хэммингу.

Keywords: the Entropy of long codes with dependent digits, regularization of calculations, a variety of Hamming convolutions, requirements for transcoding data before Hamming clipping.

Проблема вычисления энтропии длинных кодов с зависимыми разрядами

Если пытаться вычислять энтропию длинных кодов по Шеннону, то мы сталкиваемся с задачей экспоненциальной вычислительной сложности. Так, для кодов длинной 256 бит, полученных от программного генератора псевдослучайных чисел, возникает 2^{256} состояний. Произведение «Война и мир» в четырех томах Льва Толстого имеет 1640 страниц, 2000 знаков на странице дает 2^{22} знаков. Пользуясь как эталонным текстом русского языка произведением «Война и мир» по Шеннону, мы можем оценивать пароли длинной до 176 бит или 22 знака. Для оценки пароля длинной в 32 случайных знака потребуется 2^{130} произведений на русском языке размерами сопоставимыми с четырьмя томами «Войны и мира». Все оцифрованные русскоязычные источники не содержат такой объем информации. Даже если бы такой эталон русскоязычного текста существовал, его анализ на обычном современном компьютере может занять тысячи лет машинного времени.

Проблема состоит в том, что, руководствуясь Шенноном, приходится обрабатывать большие массивы данных и ждать появления редких событий. Положение меняется, если мы из пространства обычных кодов переходим в пространство расстояний Хэмминга [1–3]. Для кодов длиной 256 бит расстояний Хэмминга меняется в интервале $0 \le h \le 256$, итого 257 состояний:

$$h = 256 - \sum_{i=1}^{256} (\text{"}c_i\text{"}) \oplus (\text{"}x_i\text{"}), \tag{1}$$

где "c_i" – разряд кода длинного пароля; "x_i" – этот же разряд кода эталонного текста.

В работах [4–6] показано, что свертка Хэмминга может быть выполнена не только по модулю 2. Для того, чтобы обобщить результаты сверток и сделать их сопоставимыми, нормируем интервал, в котором могут меняться расстояния Хэмминга:

$$\tilde{h} = \frac{h}{\max(h)}. \tag{2}$$

В этом случае нормированные расстояния всех сверток Хэмминга всегда будут находиться в интервале от 0 до 1. Для примера на рис. 1 даны распределения нормированных расстояний Хэмминга для эталонных текстов на русском и английском языках.

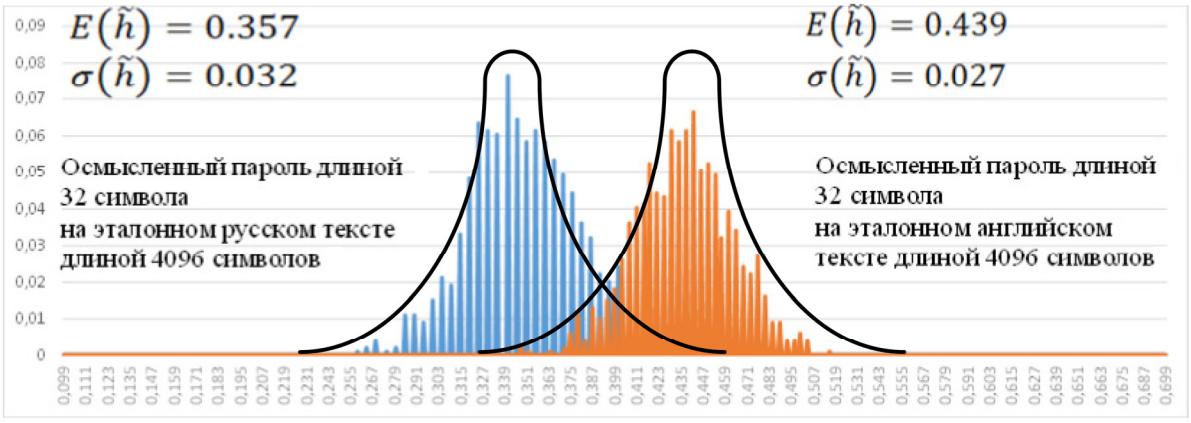

Рис. 1. Распределения расстояний Хэмминга при свертывании кода длинного осмысленного пароля с эталонными текстами на русском и английском языках

Из рис. 1 видно, что распределение расстояний Хэмминга при тестировании пароля на русском языке ближе к состоянию $\tilde{h}=0$, т.е., подбирая пароль сочетаниями фраз на русском, мы получим меньшую вероятность ошибок второго рода. В итоге оценка энтропии пароля в среде MathCAD дает величину

$$-\log\left(\text{pnorm}\left(\frac{1}{256},0.357,0.032\right),2\right) = 92.628 \text{ бит}.$$

Если мы будем пытаться осуществить атаку, подбирая пароль на русском английскими фразами, то получим очень большую оценку энтропии

$$-\log\left(\text{pnorm}\left(\frac{1}{256},0.439,0.027\right),2\right) = 192.661 \text{ бит}.$$

Смысл подобных оценок понятен, пароль на русском языке следует подбирать, пользуясь фрагментами текстов на русском языке.

Следует отметить, что приведенные выше оценки являются слишком оптимистичными. Это обусловлено тем, что при вычислениях мы не принимали в расчет 8-битную кодировку символов. Учет 8-битной структуры кодов ASCII приводит к необходимости вычислять свертки Хэмминга по модулю 8

$$h_8 = 256\cdot 32 - \sum_{i=1}^{32}\left("c_i,c_{i+1},...,c_{i+8}"\right)\oplus_8\left("x_i,x_{i+1},...,x_{i+8}"\right). \quad (3)$$

В итоге мы получаем более реалистичные распределения расстояний Хэмминга, приведенные на рис. 2.

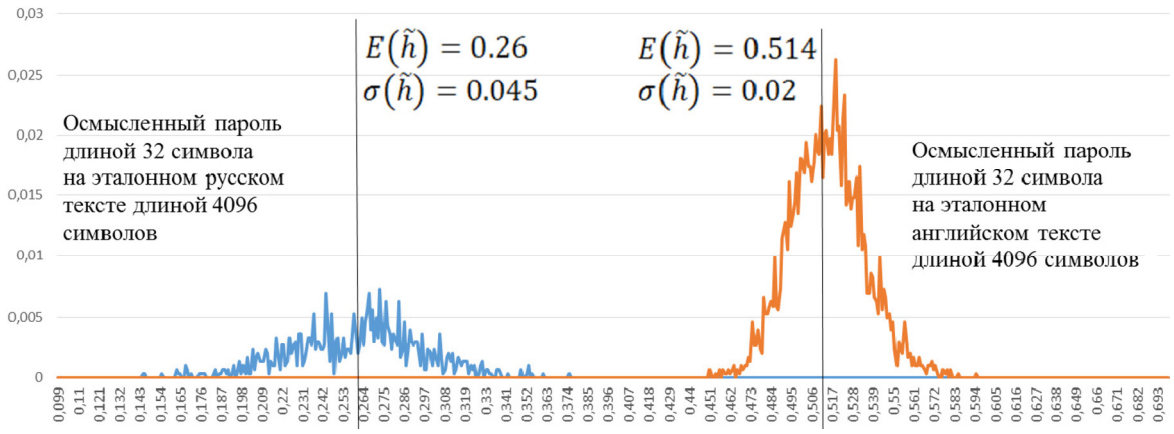

Рис. 2. Распределение расстояний Хэмминга в 8-битной системе счисления со свертыванием данных по модулю 8

Оценка энтропии для осмысленного пароля на русском при его тестировании тоже на русском получается ниже:

$$-\log\left(\text{pnorm}\left(\frac{1}{256\cdot 32},0.26,0.045\right),2\right) = 27.954 \text{ бит}.$$

Если мы такую же оценку выполняем применяя сочетания слов на английском, то получаем увеличение энтропии:

$$-\log\left(\text{pnorm}\left(\frac{1}{256\cdot 32},0.514,0.02\right),2\right) = 482.228 \text{ бит}.$$

И в том и в другом случае получаются гораздо более реалистичные оценки энтропии. И в двоичной и восьмеричной системах сверток Хэмминга мы наблюдаем дефект вычислений (неустойчивость метода), когда тестируем пароль на другом языке. В восьмеричной системе счисления этот дефект усилился.

Наличие этого дефекта связано с тем, что ASCII кодировки имеют компактное расположение кодов букв латиницы и кодов букв кириллицы. Оба этих алфавита имеют расстояние между центрами групп «латиницы» и «кириллицы» 224 – 96 = 128 (7 бит). Именно это обстоятельство и приводит к расхождению математических ожиданий расстояний Хэмминга распределений (см. рис. 1 и 2).

На величину стандартного отклонения распределения расстояний Хэмминга прежде всего влияет компактность кодировки групп символов (отсутствие разрывов между кодами). Как следствие, сделать процедуры вычисления сверток Хэмминга более устойчивыми удается перекодировками, которые ликвидируют пробелы между кодами в группах «латиница» для текстов на английском и «кириллица» для текстов на русском. Часто используемые в текстах знаки препинания должны иметь коды в группе символов в соответствии с вероятностью их появления в тексте. Группировка кодов и их упорядочивание по частоте появления символов являются мощными методами структурной регуляризации вычислений энтропии. Один из возможных примеров данных методов регуляризации вычисления энтропии является кодировка, приведенная в табл. 1.

Таблица 1

Таблица перекодировки групп символов «кириллица» для текстов на русском языке для регуляризации вычислений энтропии

Символ	Код символа	Символ	Код символа	Символ	Код символа	Символ	Код символа
	0	ь	18	О	36	ъ	54
о	1	ы	19	К	37	Р	55
е	2	г	20	Л	38	Г	56
а	3	б	21	С	39	У	57
н	4	ч	22	Д	40	З	58
и	5	з	23	-	41	Ф	59
т	6	.	24	И	42	Х	60
с	7	ж	25	П	43	Ш	61
л	8	й	26	Я	44	Щ	62
в	9	ш	27	ф	45	Ж	63
р	10	х	28	Т	46	Ц	64
к	11	ю	29	М	47	«	65
,	12	э	30	…	48	Ь	66
д	13	А	31	:	49	Ю	67
м	14	щ	32	Ч	50	е	68
у	15	ц	33	Е	51	Й	69
п	16	В	34	Э	52	Ъ	70
я	17	Н	35	Б	53	Ы	71

На рис. 3 сопоставлены результаты вычисления распределения расстояний Хэмминга для нескольких осмысленных паролей на русском языке длиной 256 бит, в кодировке ASCII и оптимальной кодировке в соответствии с табл. 1.

Из рис. 3 видно, что вычисление распределения расстояний Хэмминга при тестировании паролей на русском языке, представленных в оптимальном коде, имеет ряд преимуществ перед вычислением расстояний Хэмминга этих же паролей, представленных в кодировке ASCII:

1. После замены кодировок исчезла мультимодальность распределения расстояний Хэмминга, что делает гипотезу нормальности данных корректной.

2. Видна методическая ошибка по вычислению математического ожидания в кодировке ASCII, которая устраняется в новой кодировке (математическое ожидание уменьшается, соответственно должна снижаться энтропия, что приближает оценки к классическим по Шеннону).

3. Примерно в 2 раза снижается стандартное отклонение, что эквивалентно значительному повышению устойчивости вычислений.

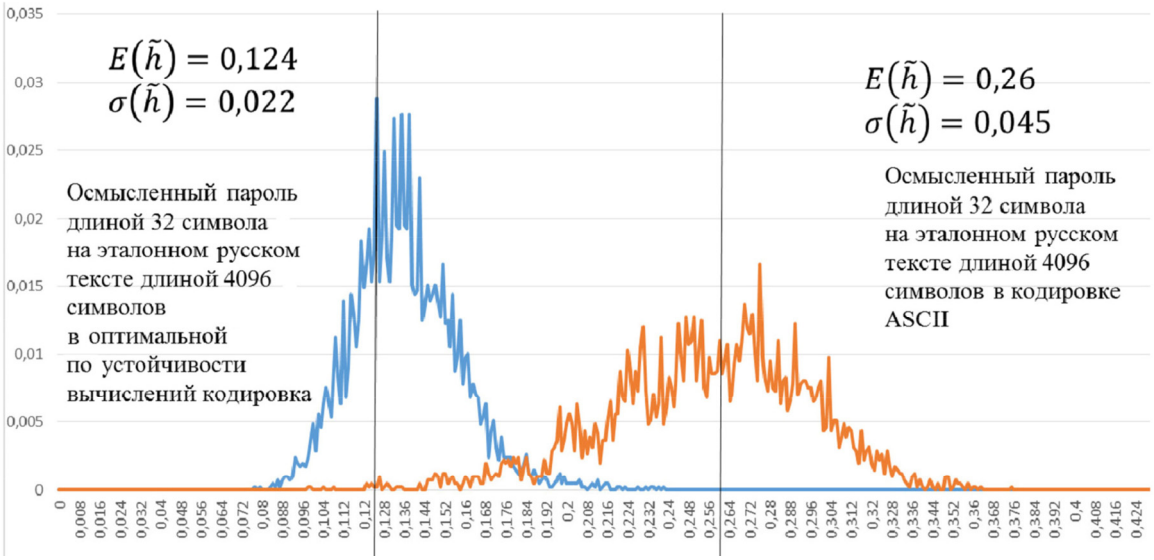

Рис. 3. Соотношение распределений расстояний Хэмминга в оптимальном коде и кодировке ASCII

Таким образом, оценку энтропии в пространстве сверток Хэмминга можно сделать еще более устойчивой, если осуществлять предварительную перекодировку символов ASCII по специальной кодировке, обеспечивающей минимизацию значения математического ожидания расстояний Хэмминга и их стандартного отклонения.

Библиографический список

1. *Иванов, А. И.* Оценка усиления стойкости коротких цифровых паролей (PIN кодов) при их рукописном воспроизведении / А. И. Иванов, О. В. Ефимов, В. А. Фунтиков // Защита информации. INSIDE. – 2006. – № 1. – С. 55–57.
2. *Малыгин, А. Ю.* Быстрые алгоритмы тестирования нейросетевых механизмов биометрико-криптографической защиты информации / А. Ю. Малыгин, В. И. Волчихин, А. И. Иванов, В. А. Фунтиков. – Пенза : Изд-во ПГУ, 2006. – 161 с.
3. ГОСТ Р 52633.3–2011. Защита информации. Техника защиты информации. Тестирование стойкости средств высоконадежной биометрической защиты к атакам подбора.
4. *Юнин, А. П.* Оценка энтропии легко запоминаемых, длинных паролей со смыслом в ASCII кодировке для русского и английского языков / А. П. Юнин, О. В. Корнеев // Тестирование стойкости средств высоконадежной биометрической защиты к атакам подбора : тр. науч.-техн. конф. кластера пензенских предприятий, обеспечивающих безопасность информационных технологий. – Пенза, 2016. – Т. 10. – С. 40–42. – URL: http://пниэи.рф/activity/science/BIT/T10-p40.pdf
5. *Волчихин, В. И.* Многомерный портрет цифровых последовательностей идеального «белого шума» в свертках Хэмминга / В. И. Волчихин, А. И. Иванов, А. П. Юнин, Е. А. Малыгина // Известия высших учебных заведений. Поволжский регион. Технические науки. – 2017. – № 4. – С. 4–13.
6. *Иванов, А. И.* Многомерная нейросетевая обработка биометрических данных с программным воспроизведением эффектов квантовой суперпозиции / А. И. Иванов. – Пенза : Изд-во АО «ПНИЭИ», 2016. – 133 с. – URL: http://пниэи.рф/activity/science/BOOK16.pdf

References

1. Ivanov A. I., Efimov O. V., Funtikov V. A. *Zashchita informatsii. INSIDE* [Information protection. INSIDE]. 2006, no. 1, pp. 55–57. [In Russian]
2. Malygin A. Yu., Volchikhin V. I., Ivanov A. I., Funtikov V. A. *Bystrye algoritmy testirovaniya neyrosetevykh mekhanizmov biometriko-kriptograficheskoy zashchity informatsii* [Fast algorithms for testing neural network mechanisms of biometric and cryptographic protection of information]. Penza: Izd-vo PGU, 2006, 161 p. [In Russian]
3. GOST R 52633.3–2011. *Zashchita informatsii. Tekhnika zashchity informatsii. Testirovanie stoykosti sredstv vysokonadezhnoy biometricheskoy zashchity k atakam podbora* [GOST R 52633.3–2011. Information protection. Information security techniques. Testing resistance means highly reliable biometric security to attacks selection]. [In Russian]

4. Yunin A. P., Korneev O. V. *Testirovanie stoykosti sredstv vysokonadezhnoy biometricheskoy zashchity k atakam podbora: tr. nauch.-tekhn. konf. klastera penzenskikh predpriyatiy, obespechivayushchikh bezopasnost' informatsionnykh tekhnologiy* [Testing the resistance of highly reliable biometric protection to attacks of selection: tr. scientific.-tekhn. conf. cluster of Penza enterprises providing information technology security]. Penza, 2016, vol. 10, pp. 40–42. Available at: http://pniei.rf/activity/science/BIT/T10-p40.pdf [In Russian]
5. Volchikhin V. I., Ivanov A. I., Yunin A. P., Malygina E. A. *Izvestiya vysshikh uchebnykh zavedeniy. Povolzhskiy region. Tekhnicheskie nauki* [University proceedings. Volga region. Engineering sciences]. 2017, no. 4, pp. 4–13. [In Russian]
6. Ivanov A. I. *Mnogomernaya neyrosetevaya obrabotka biometricheskikh dannykh s programmnym vosproizvedeniem effektov kvantovoy superpozitsii* [Multidimensional neural network processing of biometric data with software reproduction of quantum superposition effects]. Penza: Izd-vo AO «PNIEI», 2016, 133 p. Available at: http://pniei.pf/activity/science/BOOK16.pdf [In Russian]

Волчихин Владимир Иванович
доктор технических наук, профессор,
президент Пензенского государственного университета
(Россия, г. Пенза, ул. Красная, 40)
E-mail: president@pnzgu.ru

Volchikhin Vladimir Ivanovich
doctor of technical sciences, professor,
President of Penza State University
(40 Krasnaya street, Penza, Russia)

Иванов Александр Иванович
доктор технических наук, профессор,
начальник лаборатории,
Пензенский научно-исследовательский электротехнический институт
(Россия, г. Пенза, ул. Советская, 9)
E-mail: pniei@penza.ru

Ivanov Aleksandr Ivanovich
doctor of technical sciences, professor,
head of the laboratory,
Penza Scientific Research Electrotechnical Institute
(9 Sovetskaya street, Penza, Russia)

Карпов Артем Павлович
аспирант,
Пензенский государственный университет
(Россия, г. Пенза, ул. Красная, 40);
E-mail: artem.karpei@mail.ru

Karpov Artem Pavlovich
postgraduate student,
Penza State University
(40 Krasnaya street, Penza, Russia)

Юнин Алексей Петрович
специалист,
Пензенский научно-исследовательский электротехнический институт
(Россия, г. Пенза, ул. Советская, 9)
E-mail: pniei@penza.ru

Yunin Alexey Petrovich
specialist,
Penza Scientific Research Electrotechnical Institute
(9 Sovetskaya street, Penza, Russia)

Образец цитирования:

Условия корректного вычисления энтропии осмысленных длинных паролей в пространстве сверток Хэмминга с эталонными текстами на русском и английском языках / В. И. Волчихин, А. И. Иванов, А. П. Карпов, А. П. Юнин // Измерение. Мониторинг. Управление. Контроль. – 2019. – № 3 (29). – С. 33–38. – DOI 10.21685/2307-5538-2019-3-4.