УДК 519.2, 004.622, 004.056.53 DOI 10.21685/2307-5538-2019-4-5

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

ВОЗМОЖНОСТИ РЕГУЛЯРИЗАЦИИ ВЫЧИСЛЕНИЙ ЭНТРОПИИ ДЛИННЫХ ОСМЫСЛЕННЫХ ПАРОЛЕЙ В ПРОСТРАНСТВЕ СВЕРТОК ХЭММИНГА

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

POSSIBILITIES IF REGULARIZING THE ENTROPY CALCULATIONS OF LONG MEANINGFUL PASSWORDS IN HAMMING CONVOLUTIONS SPACE

Аннотация. ***Актуальность и цели.*** Целью работы является исследование возможностей регуляризации вычисления энтропии длинных кодов с зависимыми разрядами, являющимися осмысленными легко запоминаемыми паролями на родном языке пользователя. ***Материалы и методы.*** Используется пространство сверток Хэмминга кодовой последовательности парольной фразы длиной в 256 бит с эталонной кодовой последовательностью длиной 32 000 бит. Статистическими методами исследуются различные методы повышения точности вычисления энтропии длинных осмысленных паролей в пространстве сверток Хэмминга. ***Результаты.*** Показано, что стандартное отклонение распределения значений расстояний Хэмминга падает при отказе от стандартной кодировки ASCII парольной фразы и эталонной последовательности и переходе к кодировкам с определенными свойствами. Предложено использовать кодировку без разрывов между кодируемыми символами, упорядоченную по частоте встречи символа в тексте на родном языке пользователя. ***Выводы.*** В пространстве сверток Хэмминга за счет перехода к оптимальной кодировке символов парольной фразы и эталонного текста возможно снижение требований к длине эталонного текста.

Abstract. ***Background.*** The aim of the work is to study the possibilities of regularizing the calculation of the long codes entropy with dependent digits, which are meaningful easily remembered passwords in the user's native language. ***Materials and methods.*** The Hamming convolutions space is used for a code sequence of a passphrase of 256 bits with a reference code sequence of 32 000 bits in length. Statistical methods are used to investigate various methods for increasing the accuracy of the calculation of the long meaningful passwords entropy in the Hamming convolutions space. *Results.* It is shown that the standard deviation of the distribution of Hamming distances falls when the standard ASCII coding of the passphrase and the reference sequence is abandoned, and the transition to codings with certain properties. It is proposed to use the encoding without gaps between the characters being encoded, ordered by the frequency of the character's encounter in the text in the user's native language. ***Conclusions.*** In the Hamming convolutions space, due to the transition to the optimal encoding of the characters of the passphrase and the reference text, the requirements for the length of the reference text can be reduced.

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

K e y w o r d s: entropy of long codes with dependent digits, regularization of computations, Hamming convolutions variety, requirements for recoding data before Hamming convolutions calculation.

Проблема вычисления энтропии длинных кодов с зависимыми разрядами

В настоящее время безопасность информационных ресурсов является одной из наиболее актуальных задач сетевой инфраструктуры любой организации. При этом одним из основных методов обеспечения информационной безопасности ресурсов является использование систем разграничения доступа. В данных системах применяют различные методы аутентификации, наиболее популярным из которых является метод, основанный на знании секретной информации, примером которого является парольная защита.

Однако применение данного метода, как правило, является компромиссным – наиболее предпочтительным с точки зрения информационной безопасности является использование паролей, сформированных из случайных (псевдослучайных) символов. В то время как для конечного пользователя системы разграничения доступа запоминание и использование таких паролей неприемлемо.

Одним из возможных решений данного компромисса является использование пользователем длинных осмысленных парольных фраз – легких для запоминания, но трудных для брутфорса. Согласно обновленным рекомендациям Национального института стандартов и технологий, данное решение является наиболее предпочтительным при разработке и эксплуатации систем аутентификации [1].

Для таких систем существует задача оценки стойкости легко запоминаемых, осмысленных парольных фраз. С этой задачей легко можно справиться, оценив энтропию парольной фразы по Шеннону:

$$H("x_1, x_2, ..., x_{256}") = -\sum_{i=1}^{2^{256}} P_i \cdot \log_2(P_i), \quad (1)$$

где P_i – вероятность появления одного из 2^{256} состояний кода, тестируемой парольной фразы.

Однако данное вычисление является задачей высокой вычислительной сложности. Для реализации вычислений по формуле (1) необходимо определять вероятности появления очень редких событий. Таким образом, оценка энтропии длинных паролей по Шеннону на обычной вычислительной машине оказывается технически не выполнима даже для паролей, состоящих из восьми букв (64 бита).

Проблема оценки энтропии длинных кодов может быть решена, если придерживаться рекомендаций ГОСТ Р 52633.3 [2]. Национальный стандарт рекомендует перейти от обычного представления кодов к вычислению расстояний Хэмминга по модулю два:

$$h_2 = 256 - \sum_{i=1}^{256} ("x_i") \oplus ("T_i"), \quad (2)$$

где $\oplus$ – операция сложения по модулю два; "x_i" – разряды тестируемой последовательности; "T_i" – разряды тестовой последовательности, например, осмысленных фраз русского языка в 8-битной ASCII кодировке или случайных состояний, полученных от 256-мерного генератора «белого» шума.

При этом оценка энтропии пароля строится на предсказании редких событий (ожидание редких событий по Шеннону замещается на предсказание вероятности редких событий в пространстве расстояний Хэмминга). В свою очередь, возможность предсказаний опирается на факт нормального распределения расстояний Хэмминга для кодов с длиной более 32 бит (паролей из четырех букв в 8-битной ASCII кодировке). Это позволяет, в частности, оценивать энтропию сильно коррелированных откликов нейронной сети на биометрические образы [3–5].

При условии, что объем тестовой выборки должен примерно на порядок быть больше, чем обратная величина экспериментально оцениваемой вероятности, для оценки расстояния 256-битной последовательности к эталонному тексту по Шеннону (1) требуется 2^{256+4} случай-

ных кодов, в то время как при переходе в пространство расстояний Хэмминга для вычисления математического ожидания – $E(h)$ и стандартного отклонения – $\sigma(h)$ достаточно выборки в сто опытов ($100 \approx 2^7$). Таким образом, сокращение объемов тестовой выборки при вычислении энтропии в рассматриваемом случае может достигать величины – 2^{253}.

В работе [6] показано, что более реалистичные оценки распределения расстояний Хэмминга можно получить, приняв в расчет 8-битную кодировку символов и используя в качестве эталонного текста текст на том же языке, что и тестируемая парольная фраза. Учет 8-битной структуры кодов ASCII приводит к необходимости вычислять свертки Хэмминга по модулю восемь:

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

На рис. 1 приведено распределение расстояний Хэмминга между 32-символьной парольной фразой (256 бит) и тестовым текстом на русском языке длиной в 32 000 символов в кодировке ASCII.

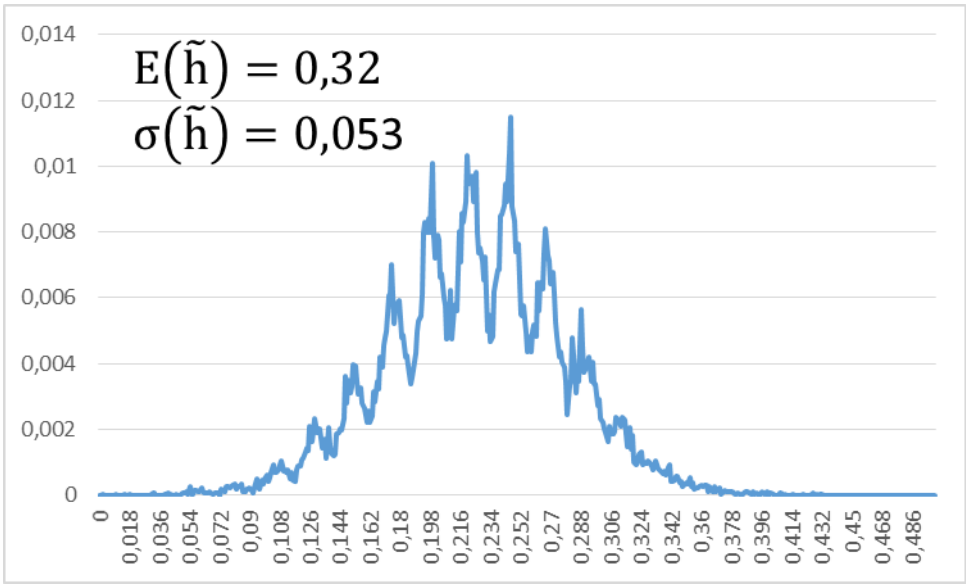

Рис. 1. Распределение расстояний Хэмминга в 8-битной системе счисления со свертыванием данных по модулю 8 в кодировке ASCII

Возможные методы регуляризации вычислений энтропии длинных кодов с зависимыми разрядами

Кодировка ASCII имеет компактное расположение кодов букв латиницы и кодов букв кириллицы. При этом в случае, если оценивать парольные фразы со знаками препинания на русском языке, можно наблюдать дефект вычислений, связанный с существенным расстоянием между группами кодов знаков препинания и кодов букв кириллицы.

Поскольку на величину стандартного отклонения распределения расстояний Хэмминга прежде всего влияет компактность кодировки групп символов (отсутствие разрывов между кодами), процедуры вычисления сверток Хэмминга можно сделать более устойчивыми за счет перекодировок, которые ликвидируют пробелы между кодами группы «кириллица» и знаков препинания для текстов на русском. Один из возможных примеров данного метода регуляризации вычисления энтропии является кодировка, приведенная в табл. 1.

На рис. 2 сопоставлены усредненные результаты вычисления распределения расстояний Хэмминга для 10 осмысленных паролей на русском языке, длиной 256 бит с эталонным текстом длиной 32 000 символов, в кодировке ASCII и кодировке без разрывов между группами символов в соответствии с табл. 1.

Таблица 1

Таблица перекодировки групп символов «кириллица»
и знаков препинания для текстов на русском языке

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

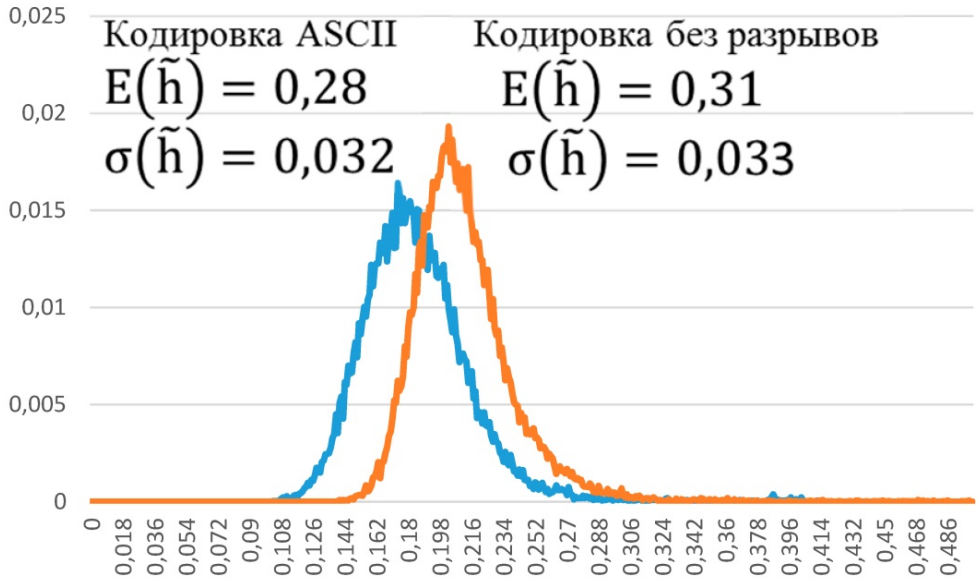

Рис. 2. Соотношение распределений расстояний Хэмминга в кодировке
без разрывов между группами символов и кодировке ASCII

Как видно из рис. 2, распределения расстояний Хэмминга имеют сопоставимые значения среднеквадратичного отклонения и математического ожидания. Это связано с тем, что отношение количества «некомпактно» расположенных кодов знаков препинания кодировки к «компактно» расположенным кодам кириллицы в ASCII-кодировке невелико – порядка 10 %.

Рассмотренные процедуры вычисления сверток Хэмминга можно сделать более устойчивыми за счет упорядочивания кодов по вероятности встречи символа в тексте. Один из возможных примеров данного метода регуляризации вычисления энтропии является кодировка, приведенная в табл. 2.

Таблица 2

Таблица перекодировки упорядоченных групп символов «кириллица» и знаков препинания для текстов на русском языке

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 усредненные результаты вычисления распределения расстояний Хэмминга для 10 осмысленных паролей на русском языке, длиной 256 бит с эталонным текстом длинной 32 000 символов, в кодировке ASCII и кодировке без разрывов между группами символов в соответствии с табл. 2.

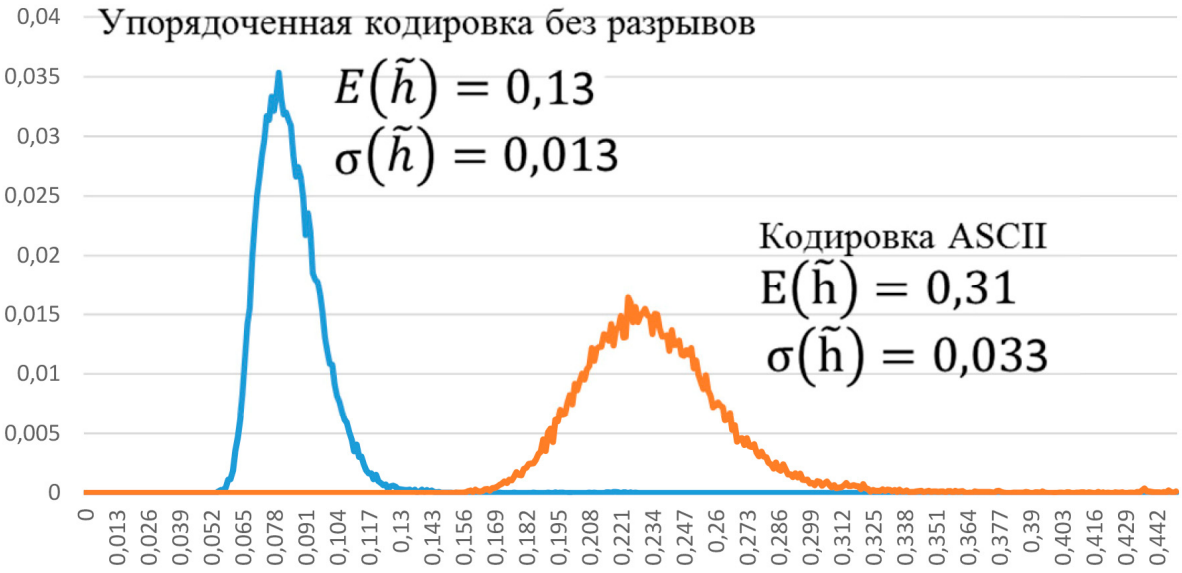

Рис. 3. Соотношение распределений расстояний Хэмминга в упорядоченной кодировке без разрывов между группами символов и кодировке ASCII

Для оценки полученных результатов на рис. 4 приведены результаты вычисления распределения расстояний Хэмминга в случайно упорядоченной кодировке без разрывов.

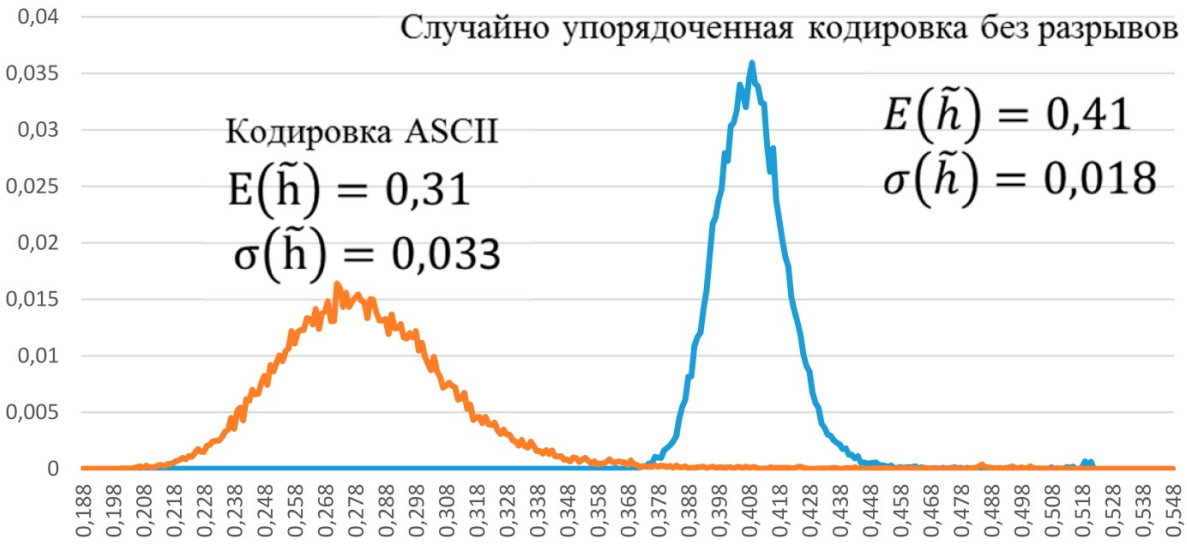

Рис. 4. Соотношение распределений расстояний Хэмминга в случайно упорядоченной кодировке без разрывов между группами символов и кодировке ASCII

Преимущества использования регуляризации вычислений энтропии длинных кодов с зависимыми разрядами

Статистические характеристики распределений расстояний Хэмминга для каждой из вышеуказанных кодировок были сведены в табл. 3.

Таблица 3

Статистические характеристики распределений расстояний Хэмминга для различных кодировок

Характеристика	ASCII	Устранены разрывы	Сортировка по возрастанию	Случайная сортировка
$\sigma(\tilde{h})$	0,032	0,033	0,013	0,021
$E(\tilde{h})$	0,28	0,31	0,13	0,41

Наибольшее среднеквадратичное отклонение (σ = 0,032) наблюдаем при использовании ASCII-кодировки, практически сопоставимое (σ = 0,033) – при устранении разрывов между кодами букв «кириллицы» и кодами знаков препинания, наименьшее (σ = 0,013) – при кодировке, без разрывов между кодами букв «кириллицы» и кодами знаков препинания с упорядочиванием кодов по вероятности встречи символа в тексте. Необходимо отметить, что среднеквадратичное отклонение при использовании случайной кодировки (без разрывов) также ниже относительно стандартной кодировки ASCII.

Руководствуясь полученными результатами, можно заключить, что при использовании «оптимальной» кодировки можно обеспечить снижение требований к тестовой выборке символов.

Таким образом, требования к тестовой выборке можно снизить, оценивая энтропию в пространстве сверток Хэмминга более устойчивыми методами, осуществляя предварительную перекодировку символов ASCII, обеспечивающую минимизацию значения математического ожидания расстояний Хэмминга и их стандартного отклонения.

Библиографический список

1. NIST Special Publication 800-63B. Digital Identity Guidelines. Authentication and Lifecycle Management. Paul A. Grassi, James L. Fenton.
2. ГОСТ Р 52633.3–2011. Защита информации. Техника защиты информации. Тестирование стойкости средств высоконадежной биометрической защиты к атакам подбора.
3. *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.

4. *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.
5. *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
6. *Волчихин, В. И.* 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 // Безопасность информационных технологий : сб. ст. по результатам Всерос. науч.-техн. конф., посвящ. 100-летию со дня рождения одного из основоположников советской вычислительной техники Б. И. Рамеева. – Пенза, 2019. – С. 59–65. – URL: http://rosoperator.ru/rameev100

References

1. *NIST Special Publication 800-63B. Digital Identity Guidelines. Authentication and Lifecycle Management*. Paul A. Grassi, James L. Fenton.
2. 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 the resistance of highly reliable biometric security tools to selection attacks]. [In Russian]
3. Ivanov A. I., Efimov O. V., Funtikov V. A. *Zashchita informatsii. INSIDE* [Information protection. INSIDE]. 2006, no. 1, pp. 55–57. [In Russian]
4. 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 for biometrico-cryptographic information protection]. Penza: Izd-vo PGU, 2006, 161 p. [In Russian]
5. 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]
6. Volchikhin V. I., Ivanov A. I., Karpov A. P., Yunin A. P. *Bezopasnost' informatsionnykh tekhnologiy: sb. st. po rezul'tatam Vseros. nauch.-tekhn. konf., posvyashch. 100-letiyu so dnya rozhdeniya odnogo iz osnovopolozhnikov sovetskoy vychislitel'noy tekhniki B. I. Rameeva* [Information technology security: a collection of articles based on the results of the all-Russian scientific and technical conference dedicated to the 100th anniversary of the birth of one of the founders of Soviet computer technology, B. I. Rameyev]. Penza, 2019, pp. 59–65. Available at: http://rosoperator.ru/rameev100 [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:

Волчихин, В. И. Возможности регуляризации вычислений энтропии длинных осмысленных паролей в пространстве сверток Хэмминга / В. И. Волчихин, А. И. Иванов, А. П. Карпов, А. П. Юнин // Измерение. Мониторинг. Управление. Контроль. – 2019. – № 4 (30). – С. 43–50. – DOI 10.21685/2307-5538-2019-4-5.