DOI: 10.15439/2021F80
ISSN 2300-5963 ACSIS, Vol. 26

On new stream algorithms generating sensitive digests of computer files

Vasyl Ustimenko
University of Marie Curie-Sklodowska in Lublin, ul.
Plac Marii Curie-Skłodowskiej 5, Lublin, 20-031,
Poland
Email: vasulustimenko@yahoo.pl

Oleksandr Pustovit
Institute of Telecommunications and the Global Informatio
the National Academy of Sciences of Ukraine, Chokolivsky
13, Kyiv, 02000, Ukraine
Email: pustovitoleksandr0709@gmail.com

The paper is dedicated to construction of new fast and flexible hash-based message authentication codes (HMACs) that will provide large files with cryptographically stable digestions in the Postquantum era.

These instruments can be used for detecting cyber-terrorist attacks, file audits and checking the integrity of messages during communication, We use algebraic properties of well known extremal graphs *D(n, q)* and *A(n,q)* with good expansion property for the construction of HMACS.

I. Introduction

We propose new fast algorithms for the creation of sensitive digests of electronic files to detect cyberattacks, computer viruses or other damages and check data integrirty. These tools can be used to defend virtual organization and conduct the audit of all files after a registered intervention. Cryptographic stability of new key-dependent hash functions is associated with complex algebraic problems, such as the study of systems of algebraic equations of large degree and the problem of decomposition of nonlinear transformation into the composition of given generators. These facts justify resistance of digests against adversary attacks with the usage of algorithms in terms of Turing machine or Quantum Computation Theory.

Algorithms of digests generation use idea of presentation of files in the form of sequences (words) of elements of finite commutative ring K , such as finite field or arithmetical ring modulo 2^m. Such words can be treated as elements of free semigroup or its modification S with the usage of additive ring operation. The presentation of k -regular tree (or other regular infinite graph) in the form of projective limit of finite graphs given by equations allows to define «compression homorphism» of S into group of polynomial transformations of affine space K^t (space of tigests of the chosen dimension (t). Affine transformations are used to hide the homomorphic map. This scheme is new.

Implemented accordingly to this scheme family of fast algorithms was investigated via computer simulations on the real data of large size. Change of single character of the document in binary alphabet causes the change of the majority of characters of produced digest (≥98%). This property and evaluation of time execution of software programs justify potential of practical usage of implemented algorithm for cybersecurity tasks.

II. On the verification of electronic documents

Protection of large data repositories against cyberattacks via creation of digests of both encoded and original electronic documents in the selected starting time is the impotent task. With a change of time new digests could be created and compared to the original ones. The presence of any changes indicates damage to the files (cyberattack, computer virus, hardware failure, staff error and more).

For the checking the integrity of electronic documents within transmission the correspondent creates a digest of the original file and the same file in an encrypted form. His/her partner creates a digest of the received decrypted document and the original encrypted document. Correspondents compare the digests and conclude whether or not document was damaged.

A simplified model of the global information space can be imagined as a large, growing network of registered virtual users (individuals or institutions) who exchange information and can store it in electronic repositories located on the network or isolated from them.

The size of files for sharing (electronic documents) tends to grow. An important category of information space is trustworthiness of documents.

Users can use a symmetric private key algorithm to encrypt documents and key exchange protocol to maintain encryption security. Certified public key algorithms may be also used to change the key. These methods ensure the security of the exchange channels.

It is easy to see that even if a reliable encryption is used it does not provide a complete trustworthines of the documents, because it is necessary to take into account the noise in the channels and the problems of safe storage of files in electronic repositories, where documents can be tampered with, damaged by computer viruses, technical errors in the work of computing machinery, etc.

It should be noted that the threat of powerful cyberterrorist attacks on repositories of electronic information are recently increasing. There consequences are not only information leakage, but also a damage or a falsification of documents. It is clear that once a cyberattack is detected on a corporate information repository you need to audit all system files. Countering these threat require the development of a new software.

Other information security tasks require a general hash function that does not require a key or password (see for instance [1], [2] and further references).

III. REQUIREMENTS TO DIGESTS

The cryptographically stable hash function f must provide the practical impossibility of selecting a pair of links x and z with the same value of the hash function. The digest of a document created with a key-dependent hash function (MAC) uses the HMAC symbol. When users want to exchange correspondence securily, verifying who is the actual author of the letter, and the absence of changes when forwarding, they choose a shared MAC. Additionally they use a common symmetric encryption scheme.

In addition to cryptographical stability the execution speed and high indicator of avalanch effect are important. Avalanch effect can be measured in the following way. The HMAC of generated file has to be computed, after this step some chosen character of original file has to be changed for other symbol and HMAC for the new file has to be computed.

Finally bit to bit comparison of characters of two HMACs has to be done and persantage of changed characters has to be computed. For practical usage of HMAC is necessary to show that change of arbitrarily used character leads to the change of at least 40 persent of bites independently from the size of tested files.

Introduced approach of usage of special subgroups of endomorphisms from CSn(K) is useful for the development of stream ciphers of Symmetric Cryptography (see for instance [3], [4], [5] and further references) and constructions of HMACs (see [6] where special linear groups have been used). We use nonlinear subgroups of $CS_n(K)$. The method of generation of nonlinear transformations of free modules over commutative rings described in the terms of special graphs defined by algebraic equations (so called linguistic graphs) can be used instead of methods of generators and equations. Other applications of graph theory to Cryptography are considered in [7].

Studies of message authentification codes and HMACs is a hot topic.Complete list of all published papers within this direction is impossible to make, we only refer to some recent papers [8] - [17].

Recall that noncommutative cryptography is an active field of cryptology that explores cryptographic primitives and systems based on algebraic structures such as groups, semigroups, and non-commutative rings.

One of the earliest applications of noncommutative algebraic structure for cryptographic purposes was the usage of groups for the development of cryptographic protocols.

The method of usage of platform G which is a subgoup or subsemigroup of affine Cremona semigroup $CS(K^n)$ defined over finite commutative ring K under the condition that each element is presented in [19]. This is an attempt to merge methods of noncommutative cryptography and multivariate cryptography.

Studies of message authentification codes and HMACs is a hot topic. Noteworthy that arbitrary hash function such as MD5 or SHA-1 can be used for the composition of HMACs corresponding to MD5 and SHA-1 message authntication codes are known as HMAC-MD5 and HMAC-SHA-1 respectively. HMAC's cryptographic performance depends on the cryptographic performance of the underlying hash function, the size of its hash output, the size and quality of the key.

IV. MATHEMATICAL BACKGROUND OF PROPOSED HASH FUNCTIONS

Let $F(K)$ be a space of potentially infinite texts in alphabit K which is the totality of all tuples of kind $(a_1, a_2,..., a_k), a_i \in K$ of different case k. Assume that K is finite commutative ring and identify $F(K)$ with the semigroup with the following operation $(a_1, a_2,..., a_k) \circ (b_1, b_2,..., b_s) = (a_1, a_2,..., a_k, b_1 + a_k, b_2 + a_k,$ $b_s + a_k)$

Let $F'(K)$ be the subsemigroup of all words (tuples) of even length. We assume that CST(K n) stands for the semigroup of all polynomial maps of affine space K^n in itself.

Our algorithm is based on the following mathematical statement.

Theorem 1 (see [20]). For each natural integer $m \geq 2$ there exists homomorphism ψ: $F'(K) \rightarrow CS_m(K)$ such that its image $\psi(F'(K))$ is a group G of cubic polynomial transformations of degree 3.

Recall that the property of $\psi = \psi_m$ to be homomorphic map means that $\psi(a \circ b) = \psi(a) \circ \psi(b)$.

Transformations satisfying conditions of the theorem are defined in constructive way in terms of the theory of discrete dynamic systems defined via algebraic graphs with extremal properties. These methods allow to get the lower bound $|G| \geq 2^{4n}$ of the order of *G*. Noteworthy that the proposition defines rare mathematical object. Superposition of two randomly chosen cubic maps will have degree 9 , in the case of 3 such maps resulting degree will be 27, composition of 4 such maps has degree 81, but in the constructed group all compositions of several maps will have degree ≤3.

It was not the *G* group itself that was used to create the MAC but the mapping ψ that defines it along with the affine *A* and *B* transformations of the Cremony group with the rule $g : x \rightarrow A\psi(x)B$. It is not hard to see that it's a natural data compression operator that maps an infinite set of all even-length words in the alphabet *K* to a finite set. The output is a list of coordinates *g (x)* to which the full differential operator is applied twice. Computer simulation made it possible to calculate a very high avalanche effect within 97-98 percents. For example, in MAC of Russian researchers the avalanche effect interval is estimated as 47-50% [18]. The constructive definition of compression homomorphisms is defined in terms of the theory of linguistic graphs. The known linguistic graphs *A (n, K)* and *D (n, K)* constructed for solving some problems of extreme graph theory are used (see [21] and furtherreferences).

V. On the option to speed up the algorithm

In this unit we present the modification of described above algorithm which allows to present (or even improve) the level of riched avalanche effect under essential increase of execution time. We have to admit that algorithm is described ''by modulo'' of computation of homomorphism value in a given point. Constructive definition of ψ were already described in the previous sections.

Let $(a_1, a_2, ..., a_n)$ be digital document presented in the alphabet *K* after the merge of file with some pseudorandom word of constant length.. We assume that parameter n is even. Users select the size of digest $m, m < n$ where $m = O(1)$ or $m = O(n)$ together with the key is formed by increasing sequence of positive integers $i(1), i(2), ..., i(m-1)$ and nonsingular matrix *M* with entries from commutative ring Z_{256} of residues modulo 256. Users form vector $u = (v_1, v_2, ..., v_m)$, where $v_1 = a_1 + a_2 + ... + a_n, ..., v_j = v_{j-1} - a_{i(j-1)}$. Secondly they compute the cubical map $F = \psi_m(a_1, a_2, ..., a_n)$ and its value on the vector u. Computed row vector $F(u)$ has to be multoplied on matrix M. Vector $w = F(u)M$ is the digest of document.

Note, that the value of $F(u)$ is calculated via recursive procedure, its complexity is approximated as $O(mn)$ and coincides with the complexity of digest generation.

This basic algorithm is easy to modify without the change of computational complexity. In particular the folloving variants can be used.

One can present the word $(a_1, a_2, ..., a_n)$ in a form of concatenation of finite number of words $z_1, z_2, ..., z_t$ of even length. Secondly he/she selects the sequence $u_1, u_2, ..., u_k$ where $u_i \in \langle z_1, z_2, ..., z_l \rangle$ such that each word z_i appears at least one time in this sequence. The next step is a computation of value of product of $u_1, u_2, ..., u_k$ in the presented above semigroup of words $F'(K)$. Algorithm is modified via the change of cubic map $\psi(a)$ for $\psi(y)$. In the case of open partition of file cryptographic stability such digest rests on the decomposition problem of $\psi(y)$ into the product of transformation $\psi(z_i)$ from affine Cremona group. Noteworthy that the polynomial postquantum algorithm for solving this problem is unknown.

In fact this problem appears under the condition of the incomplete knowledge because only the value $\psi(y)$ is known but not the cubical map itself. In this modification users have to understand that the partition of a on subwords z_i and the sequence u_j are considered as a part of common private key for correspondents.

2) Correspondents can compute v_1 as aproduct of expressions $2a_i + 1$ and obtain v_i by division of v_{i-1} on $2a_{i(j-1)} + 1$.

3) In the case 2 one can change v_i for its odd powers k, k<128. Then these degrees have to be counted as parameters of private key.

Implemented cases are convenient for their usage in blockchain technologies where digests in the form of sequence of bites 0 and 1 symbols are needed.

We have to note that good mixing properties of compression maps are based on the constructions of homomorphisms of infinite semigroup of words of even length in affine Cremona group defined via families of algebraic graphs with remarkable extremal properties.

VI. On the implementation of digest generation algorithms

Programs are inmplemented in C++ language. Time execution of a software depends on the parameters of a computer. We use ordinary personal computer with

Pentium 3.00 GHz processor, 2GB of RAM memory for Windows 7 system.

For the computer simulations with presented above basic algorithm on the base of group *GA(n,K)* we use sparse matrix M, computable in time *O(m)* where m is digest size.

Digests were presented in characters of binary alphabet to measure of avalanch effect. Time execution in seconds for files of various size is presented below.

Table 1 – Time execution of digests generation

Size of file (megabytes)	Size of digests (in bites)		
	256	512	1024
4,0	1,36	2,74	5,52
16,1	4,94	9,90	19,82
38,7	11,60	23,20	46,46
62,3	18,54	37,10	74,22
121,3	36,24	72,52	145,02
174,2	51,22	103,66	207,34

Computer simulation demonstrates that the change of a single character of an electronic document leads to the change of 98% of the corresponding digest.

VII. Conclusion

The routine work of an enterprise, corporation, financial institution requires a long-term work of specialists with a large number of electronic documents. Specialists must use proven information to make sound planning decisions. The validation tool for checking the documents can be large files compression algorithm producing a digest of a certain size, sensitive to any change in input characters.

New family of key-dependent fast algorithms for creating electronic documents digest is proposed. Computer simulation allows to investigate the high level of an emerging avalanche effect. Let *K* be a freely chosen finite commutative ring and *m* is a positive integer. The algorithms use the recently found homomorphic compression mapping of a semigroup of potentially infinite texts in the alphabet *K* to a finite group of cubic polynomial transformations of an affine space K^m.

The cryptographic stability of hashing functions is associated with complex algebraic problems, such as the investigation of large-scale algebraic equation systems and the problem of decomposition of a nonlinear mapping of a free module by given generators.

Algorithms are implemented in the cases of finite fields $F_2^8, F_2^{16}, F_2^{32}$, arithmetic ring Z_{256} and B(32) (Boolean ring of order 2^{32}).

Computer simulation demonstrates that the speed of the algorithm increases with the size of the base switching ring.

The proposed algorithms can handle data in the form of texts, video and audio files, movies , etc. The developed methods of creation of digests have flow character, the speed in the case of a constant size of digest depends linearly on size *n* of the file. The rise of parameter *n* increases the cryptographic stability. Block implementation is possible but not motivated, because fixed block size limits the number of variables of a system of nonlinear equations

The need for further research and technological development to create new key-dependent fast hash functions is linked to cybersecurity challenges, the growth of global information space, the expectation of a quantum computer, and the development of bitcoins technologies where we need to hash out arbitrary-sized inputs into the sequence of bits that is the digest of the so-called blockchains.

The proposed robust algorithms for creating sensitive digests of documents will now be practically used to detect cyberattacks and audit all system files after a logged-in intervention. This is the first successful attempt to implement the idea of non-commutative cryptography to create HMACs. We still believe that further work is needed to optimize the built algorithms, compare them with previously known HMACs and crypto-analytical studies.

References

[1] Oliynykov R., Gorbenko I., Kazymyrov O., Ruzhentsev V., Kuznetsov O., Gorbenko Yu., Dyrda O., Dolgov V., Pushkaryov A., Mordvinov R., Kaidalov D. Data Security. *Symmetric block transformation algorithm.* Ministry of Economical Development and Trade of Ukraine. DSTU 7624:2014. National Standard of Ukraine. Information technologies. Cryptographic. –2015.

[2] Aumasson J.Ph, Serious Cryptography: A Practical Introduction to Modern Encryption, No Starch Press. – 2017. – 312 pp.

[3] Pustovit O., Ustymenko V., Pro zastosuvannia alhebraichnoi kombinatoryky do problem koduvannia ta kryptohrafii [On the application of algebraic combinatorics to the problems of coding and cryptography] //Matematychne modeliuvannia v ekonomitsi, № 1-2. – Kyiv. – 2017. – s. 31-46.

[4] V. Ustimenko, U. Romanczuk-Polubiec, A. Wroblewska, M. Polak, E. Zhupa, On the constructions of new symmetric ciphers based on non-bijective multivariate maps of prescribed degree, Security and Communication Networks, 2019 . Volume 2019, Article ID 2137561, 15 pages

[5] V. Ustimenko, U. Roman'czuk-Polubiec, A. Wroblewska, M. Polak and E. Zhupa, *On the implementation of new symmetric ciphers based on non-bijective multivariate maps,* Proceedings of the 2018 Federated Conference on Computer Science and Informatics.
Proceedings of the Federated Conference on Computer Science and Information Systems pp. 397–405 DOI: 10.15439/2018F204 ISSN 2300-5963 ACSIS, Vol. 15, pp.397-405.

[6] Mathew Cary, Ramarathnam Venkatesam, *A Message Authentication Code Based on Unimodular Matrix Groups*, Advances in Cryptology - CRYPTO 2003, 23rd Annual International Cryptology Conference, Santa Barbara, California, USA, August 17-21, 2003, Proceedings, Lecture Notes in Computer Science.

[7] Priyadarsini P.L.K., *A Survey on some Applications of Graph Theory in Cryptography*, Journal of Discrete Mathematical Sciences and Cryptography, 18:3, 209-217 (2015).

[8] Mihir Bellare, Daniel J. Bernstein, and Stefano Tessaro, *Hash-function based PRFs:AMAC and its multi-user security,* LNCS, pages 566-595. Springer, Heidelberg, 2016.

[9] Kan Yasuda. A Double-Piped Mode of Operation for MACs, PRFs and PROs: *Security beyond the Birthday Barrier.* In Antoine Joux, editor, EUROCRYPT, volume 5479 of Lecture Notes in Computer Science, pages 242-259. Springer, 2009.

[10] Xiaoyun Wang, Hongbo Yu,WeiWang, Haina Zhang, and Tao Zhan. *Cryptanalysis on HMAC/NMACMD5 and MD5-MAC*. In Antoine Joux, editor, EUROCRYPT, volume 5479 of Lecture Notes in Computer Science, pages 121-133. Springer, 2009.

[11] Gaetan Leurent, Thomas Peyrin, and Lei Wang. *New Generic Attacks against Hash-Based MACs.* In Kazue Sako and Palash Sarkar, editors, Advances in Cryptology-ASIACRYPT 2013-1 volume 8270, pages 11-20. 2013.

[12] Neal Koblitz and Alfred Menezes. Another look at HMAC. Cryptology ePrint Archive, Report 2012/074, 2012.

[13] Yevgeniy Dodis, Eike Kiltz, Krzysztof Pietrzak, and Daniel Wichs. Message authentication, revisited.In David Pointcheval and Thomas Johansson, editors, EUROCRYPT 2012, volume 7237 of LNCS, pages 355-374. Springer, Heidelberg, April 2012

[14] Yevgeniy Dodis and John P. Steinberger, Domain Extension for MACs Beyond the Birthday Barrier, In Kenneth G. Paterson, editor, EUROCRYPT, volume 6632 of Lecture Notes in Computer Science,pages 323-342. Springer, 2011.

[15] Yevgeniy Dodis, Thomas Ristenpart, John P. Steinberger, and Stefano Tessaro. To Hash or Not to Hash Again? ,(In) Difererentiability Results for H2 and HMAC. In Reihaneh Safavi-Naini and Ran Canetti,editors, CRYPTO, volume 7417 of Lecture Notes in Computer Science, pages 348-366. Springer, 2012.

[16] Pierre-Alain Fouque, Gaetan Leurent, and Phong Q. Nguyen. Full Key-Recovery Attacks on HMAC/NMAC-MD4 and NMAC-MD5, In Alfred Menezes, editor, CRYPTO, volume 4622 of Lecture Notes in Computer Science, pages 13-30. Springer, 2007.

[17] Jongsung Kim, Alex Biryukov, Bart Preneel, and Seokhie Hong. On the Security of HMAC and NMAC Based on HAVAL, MD4, MD5, SHA-0 and SHA-1 (Extended Abstract). In Roberto De Prisco and Moti Yung, editors, SCN, volume 4116 of Lecture Notes in Computer Science. Springer, 2006.

[18] Krendelev S., Sazonova P., *Parametric Hash Function Resistant to Attack by Quantum Computer*, Based on Problem of Solving a System of Polynomial Equations in Integers, Proceedings of the 2018 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS. – Vol. 15. –pp. 387 – 390 (2018)

[19] V. A. Ustimenko, On the Families of Stable Multivariate Transformations of Large Order and Their Cryptographical Applications, Tatra Mountains Mathematical Publications,2O17, 70(1), pp 107-117.

[20] V. A. Ustimenko, On multivariate public keys based on the pair of transformations with the density gap. Доповіді НАН У, 2018. 9, c. 21-27.

[21] V.Ustimenko, On the usage of postquantum protocols defined in terms of transformation semi-groups and their homomorphisma, Theoretical and Applied Cybersecurity, National Technical University of Ukraine "Igor Sikorsky Kiev Polytechnic Institute", vol 2, 2020, pp. 32-44.