The ISC Int'l Journal of Information Security
ISeCure

November 2020, Volume 12, Number 3 (pp. 55–58)
http://www.isecure-journal.org

A New Derivation of the Leftover Hash Lemma**

Marcelo S. de Alencar *,1 and Karcius D. R. Assis 1
1 *Institute of Advanced Studies in Communications, Federal University of Bahia.*

ARTICLE INFO.

Keywords:
Hash Lemma, Coding, Probability

Type: Research Article

doi: 10.22042/isecure.2021.271070.622

ABSTRACT

This paper reviews the characteristics of the main digest algorithms, and presents a new derivation of the leftover hash lemma, using the collision probability to derive an upper bound on the statistical distance between the key and seed joint probability, and the hash bit sequence distribution.

1 Introduction

A hash function is a function that maps a data file of arbitrary size onto a discrete sequence. The idea was introduced in cryptography, in 1976, by Diffie and Hellman, who identified the need for a one-way hash function as a building block of a digital signature scheme, as an algorithm to protect the authenticity of information. But, it is clear that the tool is useful to solve other security problems in communication and computer networks [1].

2 The Hash Function

The values returned by a hash function are called hash values, hash codes or digests. The hash function is usually used in combination with a hash table, a data structure used for rapid data lookup. The hash function permits to speed up a database lookup by detecting duplicated records in a large file.

In mathematics, engineering, computing and cryptography, universal hashing refers to the process of selecting a hash function at random, from a family of hash functions with a certain mathematical property. This guarantees a low number of collisions in average, even if the data is chosen by an adversary.

The privacy amplification permits the extraction of secret information, probably to be used as a cryptographic key, from a large data volume, that is only partially secret. The privacy amplification allows a large set of applications, according to the key [2].

The paper next sections present a new derivation of the leftover hash lemma, based on the Rnyi entropy of order two, also known as collision probability, to derive an upper bound on the statistical distance between the key and seed joint probability, and the hash bit sequence distribution.

3 Objectives of the Hash Function

A cryptographic hash function is used to verify whether a data file maps onto a certain hash value. On the other had, it is difficult to reconstruct the information based on the hash value. Therefore, it is used to assure data integrity, and is the building block of a Hash-based Message Authentication Code (HMAC), which provide message authentication.

The ideal cryptographic hash function has some desired properties. It is a deterministic function, that is, identical messages result in the same hash. It is fast to compute the hash value for a given message. It is impracticable to generate a message from its hash value, except by trying all possible messages, therefore the hash is a one-way function.

Additionally, a small change to a message should cause an avalanche effect, that is, it changes the digest so considerably, that the new hash value seems uncorrelated with the old hash value. It is collision resistant, that is, it is impractical to find two different messages with the same hash value. A cryptographic hash function should resist attacks on its pre-image.

* Corresponding author.
**Selected Paper at the ICCMIT'20 in Athens, Greece.

Email addresses: malencar@iecom.org.br, karcius.assis@ufba.br

ISSN: 2008-2045 © 2020 ISC. All rights reserved.

The hash function creates a unidirectional process, that makes it impossible to guess the original contents of a file, based only on the message digest. In the following there is a short description of some known commercial algorithms [3].

Message-Digest algorithm 4 (MD4), an algorithm developed between 1990 and 1991 by Ron Rivest. Is suffered several attacks, and has been considered an insecure algorithm.

It is described in RFC 1320; MD5, described in RFC 1321. The Message-Digest algorithm 5 (MD5) is a hash algorithm with 128 bits, developed by RSA Data Security, Inc. and described in RFC 1321. It is used in peer-to-peer protocols (P2P) for identity verification and logins. Attach methods have been published for the MD5. The algorithm produces a digest with 128 bits (16 bytes).

The Secure Hash Algorithm 1 (SHA-1) algorithm was jointly developed by the National Institute of Standards and Technology (NIST) and by the National Security Agency (NSA), in the United States. It presented problems, and the new versions SHA-2 and SHA-3 are in use. The program produces a digest with 160 bits (20 bytes).

The Secure Hash Algorithm 3 (SHA-3) was announced by NIST, in 2015. It is a subset of the Keccak cryptographic family, developed by Guido Bertoni, Joan Daemen, Michael Peeters and Gilles Van Assche. The algorithm produces digests with 224 bits, 256 bits, 384 bits and 512 bits. In the Bitcoin blockchain, the mining process is conducted by running SHA-256 hashing functions.

Whirlpool, is a cryptographic hash function developed by Paulo S. L. M. Barreto and por Vincent Rijmen, and adopted the International Organization for Standardization (ISO) and by the International Electrotechnical Commission (IEC) as part of the ISO 10118-3 standard. The algorithm produces a digest with 512 bits (64 bytes).

The RACE Integrity Primitives Evaluation Message Digest (RIPEMD-160) represents a family of cryptographic hash functions developed, in 1996, by Hans Dobbertin, Antoon Bosselaers and Bart Preneel, from the research group COSIC, Katholieke Universiteit Leuven. The software produces a digest with 160 bits (20 bytes).

BLAKE2 was created, in 2012, by Jean-Philippe Aumasson, Samuel Neves, Zooko Wilcox-O'Hearn, and Christian Winnerlein to replace algorithms MD5 and SHA-1.

4 Mathematical Preliminaries

Consider the random variable $X \in \chi$, in a non-empty finite set, and two probability distributions, P and Q, defined in this set. Define the statistical distance, a measure of distance between the referred probability distributions, as

$$\Delta[P,Q] = \frac{1}{2}\sum_{x\in\chi}|p(x) - Q(x)|= \frac{1}{2}\sum_{l=1}^{L}|p(x_l) - q(x_l)|. \tag{1}$$

In which $L = |\chi|$ is the number of distinct symbols in the message or file, $p(x_l)$ and $q(x_l)$ represent the individual symbols probabilities, for the respective distributions P and Q. The collision probability is defined as,

$$P_C(X) = \sum_{x\in\chi} p_X^2(x) = \sum_{l=1}^{L} p^2(x_l). \tag{2}$$

The uniform distribution has a coincidence index, or collision probability $P_C(X) = \frac{1}{L}$, and any other distribution has a larger index. The Rnyi entropy of order $\alpha(\alpha > 0), \alpha \neq 1)$ is defined as [4]

$$H_\alpha(X) = \frac{1}{(1-\alpha)} \log \sum_{l=1}^{L} p^\alpha(x_l), \quad \alpha \neq 1. \tag{3}$$

The following development demonstrates and inethat involves the collision probability and the statistical distance from Formula 2, it is possible to obtain

$$\sum_{x\in\chi}\left[p_X(x) - \frac{1}{L}\right]^2 = P_C(X) - \frac{1}{L}. \tag{4}$$

Considering that $Y \in y$ is a random variable defines as

$$Y = \left[p_X(x) - \frac{1}{L}\right]. \tag{5}$$

It mean squared value is given by

$$E|Y^2|= \frac{1}{L}\left[P_C(X) - \frac{1}{L}\right] = \frac{LP_C(X) - 1}{L^2}. \tag{6}$$

Applying the inequality $E^2|Y|\leq E|Y^2|$, one obtains

$$E|Y|\leq \sqrt{E|Y^2|} = \sqrt{\frac{LP_C(X) - 1}{L^2}}. \tag{7}$$

From the definition of statistical distance, from Formula 1,

$$\Delta[P,U] = \frac{1}{2}\sum_{x\in\chi}\left|p_X(x) - \frac{1}{L}\right| = \frac{LE[Y]}{2}. \tag{8}$$

In which $U \in u$ represents the uniform distribution. Substituting the results from Equation 7, it is possible to obtain the main inequality

$$\Delta[P,U] \leq \sqrt{\frac{LP_C(X) - 1}{2}}. \tag{9}$$

That relates the statistical distance and the collision probability.

5 Privacy Amplification

The privacy amplification theorem is related to the production of universal hash functions, as described in the following [5]. First, assume that a sequence of n bits has been generated, related to the random variable X, defined in the set χ, S and Z. [6]. Consider that $\chi \in X$ is a random variable that represents the transmitted message, that S is a random variable uniformly distributed in the set of seeds S, and that $Z \in Z$ is the resulting random variable $f : \chi \ x \ S \to Z$. Then, f is called a universal hash function, if [7]

$$P|f(x,S)|= f(x), S|\leq \frac{1}{|Z|}. \tag{10}$$

In which $|\cdot|$ represents the cardinality, and the operation is valid for every choice of $x \neq x$, that are elements of χ.
In other words, any two keys of the universe collide with probability in at most $\frac{1}{|Z|}$ when the hash function f is drawn randomly from the set. This is exactly the expected probability of collision if the hash function assigned truly random hash codes to every key.
From a seed, represented by the random variable S, uniformity distributed in the set S, it is possible to generate a safe key $K = f(X,S)$, of length r, with the application of a universal hash function X and S, as follows, $f : \chi \ x \ S \to \{0,1\}^r$, in which $1 \leq r \leq \infty$. Given that the hash function f and the random variable S are public, the question is to know if the produced key is in fact safe.
The following result, known as the leftover hash lemma establishes that, given that the entropy of the sequence of n bits of X is superior to the sequence of the r bits of the key, the key is supposed to have been uniformly generated [8]:

$$\Delta\left[P_{K,S}, U_{\{0,1\}}, P_S\right] \leq \frac{1}{2} \cdot \left[2^{-\frac{|H_2(X)-r|}{2}}\right] \tag{11}$$

In which $P_{K,S}$ is the joint distribution of the key and seed, P_S is the seed distribution, $U_{\{0,1\}}$ is the hash sequence uniform distribution, of length r and $H_2(X)$ is the Rnyi entropy of order two.
This result is also known as the privacy amplification theorem. In the following, there is a new derivation of the leftover hash lemma, that uses the collision probability.
In order to obtain Equation 11, consider the following equality, obtained from the Rnyi entropy, for the case $\alpha = 2$,

$$H_2(X) = -\log P_C(X) = -\log \sum_{i=l}^{L} p^2(x_l). \tag{12}$$

This entropy is the negative of the logarithm of the coincidence index.
It follows from the Jensen inequality [9], that the Rnyi entropy of order two, or collision entropy, is limited by the Shannon entropy

$$H_2(X) \leq H(X) \tag{13}$$

with equality, if and only if, the probability distribution of X is uniform.
Substituting the adequate parameters, one obtains

$$\Delta\left[P_{K,S}, U_{\{0,1\}}, P_S\right] \leq \sqrt{\frac{|K|P_C(K,S)-1}{2}}. \tag{14}$$

Recalling that the probability distribution associated to the key is uniform, and that each key has length r, one has $|K|= 2^r$. Also, substituting result Equation 8 into Equation 14, one obtains

$$\Delta\left[P_{K,S}, U_{\{0,1\}}, P_S\right] \leq \sqrt{\frac{2^r \ 2^{-H_2(K,S)}-1}{2}}. \tag{15}$$

Considering that the mapping $f : \chi \ x \ S \to \{0,1\}^r$ cannot increase the entropy, then $H(K,S) \leq H(X)$, which, after substitution Equation 15 gives

$$\begin{aligned}\Delta\left[P_{K,S}, U_{\{0,1\}}, P_S\right] &\leq \sqrt{\frac{2^r \ 2^{-H_2(K,S)}-1}{2}} \qquad (16)\\ &\leq \sqrt{\frac{2^r \ 2^{-H_2(K,S)}}{2}}.\end{aligned}$$

recalling that the probability is always positive or null. Rewriting the inequality, exchanging the positions of the Rnyi entropy, and the number of symbols with the resulting change in the signal, one obtains the result,

$$\Delta\left[P_{K,S,}U_{\{0,1\}}, P_S\right] \leq \frac{1}{2}\left[2^{-\frac{1}{2}(K,S)-r}\right]. \tag{17}$$

6 Conclusion

The paper discussed the use of the hash function in cryptography, and presented a new derivation of the upper bound on the statistical distance between the joint distribution of the key and the seed, and the distribution of the hash bit distribution, based on the collision probability.

Acknowledgements

The authors would like to thank the Coordination for the Improvement of Higher Education Personnel (Capes), the Brazilian National Council for Scientific and Technological Development (CNPq), the Federal University of Bahia (UFBA), and the Institute of Advanced Studies in Communications (Iecom), for financing and providing the means to carry out this work.

References

[1] Deepti Bahel, Prerana Ghosh, Arundhyoti Sarkar, and Matthew A Lanham. Predicting blood donations using machine learning techniques. In *CON-*

FERENCE PROCEEDINGS BY TRACK, page 323.

[2] Bart Preneel. Cryptographic hash functions. *European Transactions on Telecommunications*, 5(4): 431–448, 1994.

[3] Charles H Bennett, Gilles Brassard, Claude Crépeau, and Ueli M Maurer. Generalized privacy amplification. *IEEE Transactions on Information Theory*, 41(6):1915–1923, 1995.

[4] Wikipedia contributors, "cryptographic hash function - wikipedia, the free encyclopedia,". [Online; acesso em 16 de julho de 2019]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Cryptographic_hash_function&oldid=905916305, 2019.

[5] Alfréd Rényi et al. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1961.

[6] Stefan Berens. *Conditional renyi entropy*. PhD thesis, Masters thesis, Mathematisch Instituut, Universiteit Leiden, 2013.

[7] J Lawrence Carter and Mark N Wegman. Universal classes of hash functions. *Journal of computer and system sciences*, 18(2):143–154, 1979.

[8] Douglas R. Stinson. Universal hashing and authentication codes. *Designs, Codes and Cryptography*, 4(3):369–380, 1994.

[9] Marcelo S Alencar and R T Alencar. *Probability Theory*. ISBN-13: 978-1-60650- 747-6 (print). New York, USA: Momentum Press, LLC, 2016.

Marcelo Sampaio de Alencar received his Bachelor Degree in Electrical Engineering, from Universidade Federal de Pernambuco (UFPE), Brazil, 1980, his master degree in electrical engineering, from Universidade Federal da Paraiba (UFPB), Brazil, 1988 and his Ph.D. from University of Waterloo, Department of Electrical and Computer Engineering, Canada, 1993. He has more than 35 years of engineering experience, and 25 years as an IEEE member, currently as senior member. For 18 years he worked for the Department of Electrical Engineering, Federal University of Paraiba, where he was full professor and supervised more than 40 graduate and several undergraduate students. Since 2003, he is chair professor at the Department of Electrical Engineering, Federal University of Campina Grande, Brazil.

Karcius D. R. Assis graduated in Electrical Engineering from the Federal University of Paraba, currently UFCG, master's degree in electrical engineering from the Federal University of Esprito Santo (UFES) and doctorate in electrical engineering from the State University of Campinas (UNICAMP). He was a postdoctoral fellow at the University of Bristol-UK from 02/2015 to 01/2016. He was a visiting fellow at the University of Essex-UK in March 2018. He was an adjunct professor at the Federal University of ABC and is currently an associate professor at the Polytechnic School of the Federal University of Bahia; Department of Electrical and Computer Engineering - reviewer of the European Journal of Operational Research (0377-2217), IEEE Communications Letters etc. He is a member of the Brazilian Computer Society (SBC) and the Brazilian Society of Optics and Photonics (SBFoton). Has experience in the area of electrical and computer engineering, with an emphasis on telecommunications systems, computer networks, computer systems and optimization; acting mainly in the following subjects: telecommunications, optical networks, network planning and optimization.