Al-Quds University

Deanship of Graduate Studies

Master of Computer Science

Thesis Approval

DIHA: Data Integrity Algorithm Using Hashing Authentication

Prepared By: Mohammed Ahmad Jamoos

Registration No: 20913510

Supervisor: Dr. Rushdi Hamamreh

Master thesis submitted and accepted, Date: Aug 20 /2013

The name and signatures of examining committee members are as follows:

1-	Head of committee	Dr.Rushdi Hamamreh	Signature: .....
2-	Internal Examiner	Dr.Nidal Al-Kafri	Signature: .....
3-	External Examiner	Dr. Aiman Abu Samra	Signature: .....

Jerusalem-Palestine

1434-2013

Abstract

The enormous development in Information Technology, storing mechanisms and transmitting data from one side to another via networks, importance of securing such information was a must.

Information security is the science that works on the protection of information and information systems against unauthorized access or modification of information, by providing the necessary tools to protect it from internal or external risks. Information security risk is recognizing procedures and protection requirements to prevent unauthorized access to information through communications, to ensure the authenticity and validity of these connections.

Information security is used since ancient times, but it began to be used effectively since the beginning of the development of information technology. **Main requirements** for information security are: Confidentiality, Authenticity, and Data Integrity.

Confidentiality refers to preventing the disclosure of information to unauthorized individuals or systems by using various encryption methods, Authenticity that ensure that information arrived from sender to the intended recipient. Data integrity means maintaining and assuring the accuracy and consistency of data to the recipient. To ensure that data is not modified or changed from unauthorized users using algorithms based on hash functions. Hash function is one way data encryption function, that cannot recover the original plain text for modification.

In this research, we have designed and developed a new One-Way Hash Algorithm. An Algorithm that is based on linear algebra concepts to generate a non-invertible matrix by using linear combination.

Our proposed algorithm was compared with MD5 and SHA-1, for the following parameters: Collisions, Time Delay, Brute force attacks and Hamming Distance .

إعداد : محمد أحمد محمود جاموس.

إشراف : الدكتور رشدي حمامرة.

ملخص

مع التطور الهائل في علم المعلومات واليات تخزينها وتناقلها بين موقع لأخر عبر الشبكات، ظهرت اهمية حفظ أمن تلك البيانات والمعلومات .

امن المعلومات يعرف بأنه العلم الذي يقوم على حماية المعلومات ونظم المعلومات من الوصول غير المصرح به او التعديل عليها ، وذلك من خلال توفير الأدوات والوسائل اللازم توفيرها لحماية المعلومات من المخاطر الداخلية او الخارجية، وكذلك المعايير والإجراءات المتخذة لمنع وصول المعلومات إلى أيدي أشخاص غير مخولين عبر الاتصالات، لضمان أصالة هذه الاتصالات وصحتها.

إن أمن المعلومات أمر قديم ولكن بدأ استخدامه بشكل فعلي منذ بدايات تطور تكنولوجيا المعلومات، ومن أهم متطلبات أمن المعلومات : السرية، التحق، وسلامة البيانات وصحتها.

أما سرية البيانات (Confidentiality) فهي لمنع الكشف عن معلومات لأشخاص غير مصرّح لهم بالإطلاع عليها أو الكشف عنها، و ذلك من خلال طرق التشفير المختلفة. التحقق (Authenticity) فهو للتأكد من وصول البيانات من المرسل إلى المتلقي المقصود. أما السلامة (Data Integrity) لضمان وصول البيانات كما أرسلت للمتلقي، وإذا تم تعديل البيانات أو تغييرها من غير مخول بذلك فهذا انتهاك لسلامة البيانات. وللحفاظ على البيانات من التغيير والتعديل من الأشخاص غير المخول لهم بذلك تستخدم الخوارزميات الهاشية (Hash Algorithm) مبنية على دالة هاشية (Hash Function)، وهي دالة تشفير ذات الاختزال تعمل بالاتجاه الواحد، وبالتالي لا يمكن الوصول الى أصل البيانات أو استرجاعه لاختراقها.

في هذه الأطروحة تم تصميم وتطوير خوارزمية مبنية على دالة هاشية، استخدم فيها الجبر الخطي (Linear Algebra) لإنتاج مصفوفات غير معكوسة (Non Invertible Matrix) باستخدام التركيبة الخطية (Linear Combination).

تم مقارنة الخوارزمية الهاشية المقترحة DILH بالخوارزميات MD5 وSHA-1 من حيث العوامل التالية: Collisions ، Time Delay ، Brute force attacks ، و Hamming Distance .

TABLE OF CONTENTS

Chapter One

1.1 Overview

In this era where most data is stored digitally, the network has already become a new lifestyle for people. This is why high security in networks is becoming a very important problem in the information age and is an important requirement[15].

The goal of cryptography is to make it possible for two people to exchange a message in such a way that other people cannot understand it. There is no end to the number of ways that it can be done, but here we will be concerned with methods of altering the text in such a way that the recipient can undo the alteration and discover the original text. Message security has been the core of cryptography and as one of cryptography's main concerns, hash can be one of the basic techniques for message security and confirming the message authenticity [1][2].

Hash algorithm is considered as the foundation algorithm of message security, identity, authentication, non-repudiation and message integrity check service. It's one of the hot spot researches in cryptography[3]. Hash algorithm is a mathematical function that converts a relatively large message into small strings of data and/or text, to check whether the message has been trampled by attackers which could compromise the information, these strings of text, called hashes, can be used [4] [5].

More precisely, a hash algorithm H maps bits strings of arbitrary finite length to string of fixed length called hashes $H(m)$ where m is the message block, given the original data, the encrypted data easy to calculate, for a given encrypted data to look for the original data is computationally infeasible, for different original data to get the same encrypted data is computationally infeasible [3][5][6].

Hash algorithm has played an important role in modern cryptography, it has generated because of the needs of message identifier. So that it is widely used in the document verification and digital signature in information security. According to its nature, you can't restore the original plaintext (P) from the encrypted cipher text (C), so it can't be used for conventional data encryption and it can only be used for identification [7][3].

MD5, SHA-1 are an excellent One-Way hash algorithm. There are three main properties of such algorithms, in addition to One-Way, it is collision-free and strongly collision-free [8][14].

Hash algorithm in practical use is mostly constituted of two parts, which could be called as compression function and iterative structure. So, we believe that the security of hash algorithm depends not only on the difficult problems, but also on the compression function structure and iterative of the hash algorithm [1][7].

There are many various hash algorithms including MD5, SHA-1, RIIPEMD and so on. However, MD5 has been broken by Wang Xiaoyun and she had found the effective method of attack the SHA-1 algorithm, so the difficulty of the break is reduced [9]. Thus, in order to meet practical application, it is desirable to study new hash generation algorithm.

1.2 Properties of Hash Functions

For a hash function to be useful for authentication it is necessary to meet these basic requirements are [1]:

1. H should accept a block of data of any size as input.
2. H should produce a fixed-length output no matter what the length of the input data is.
3. H should behave like a random function while being deterministic and efficiently reproducible.
4. H should accept an input of any length, and outputs a random string of fixed length. H should be deterministic and efficiently reproducible in that whenever the same input is given, H should always produce the same output.
5. Given a message H, it is easy to compute its corresponding digest v, meaning that v can be computed in polynomial time $O(n)$ where mv is the length of the input message, this makes hardware and software implementations cheap and practical.
6. Given a message digest v, it is computationally difficult to find m such that $H(m) = v$. This is called the One-Way or pre-image resistance property. It simply means that one should not be capable of recovering the original message from its hash value.

I. Given a message block m_1, it is computationally infeasible to find another message $m_2 \neq m_1$ with $H(m_1) = H(m_2)$. This is called the weak collision resistance or pre-image resistance property.

II. It is computationally infeasible to find any pair of distinct messages (m_1, m_2) such that $H(m_1) = H(m_2)$. This is referred to as the strong collision resistance property.

1.3 Applications of Hash Functions

The main purpose of the establishment of the hash functions is that they are used in many cryptographic protocols. However, creating and verifying digital signatures considered to be the obvious application of the hash functions in which the hash functions are responsible to maintain the authenticity of electronic documents.

However, this section is intended to provide a detailed explanation of the forms of hash functions' applications as the following:

1.3.1 Message Integrity

Hash algorithms are meant to be created to ensure the data integrity in both transmitted and stored types of data. In regard to the transmitted type of data, the principle of the hash algorithms is to guarantee that the sender and the receiver are having the exact data that had been transmitted. However, the process of verifying the data integrity of the transmitted messages passes three different stages as the following:

- The first step, is that the sender hashes the message and sends its hash's value with it. Hence, the common way of sending the hash value of it is through the use of insecure line.
- The second step, is that the receiver hashes the received message in order to ensure that the received hash value and the sent one are the same.
- The third step, is checking the integrity of the message. This step is preserved only when the two hash values are the same. If not, the integrity of the message is not guaranteed.

In regard to stored types of data, hash functions are used to preserve the data integrity. The data's hashed values are stored in one place to be compared with the new hash values in times of need.

If the previous stored hash value matches the current hash values, it is obvious that the data integrity is preserved. [1][10].

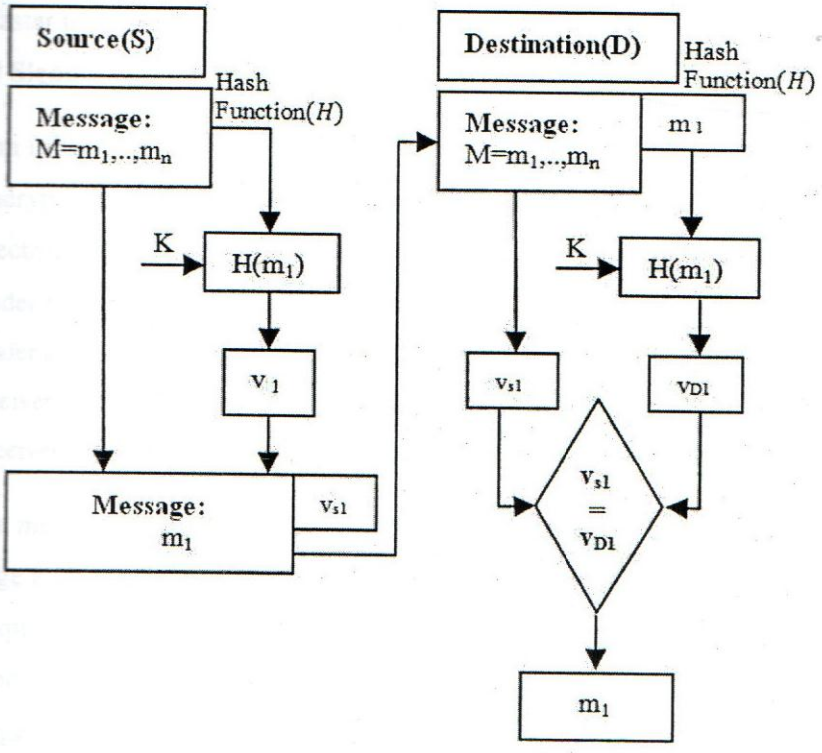

Figure 1.1: One-Way Hash Algorithm

1.3.2 Authentication Protocols

One of the protocols that are used to encrypt transmitted data is the authentication protocols. The operating principle of the authentication protocols is based on the idea of encrypted transmissions and shared secrets between communicators to decrypt the encryption. Indeed, while we are using a public key encryption, we might be subjected to a slow process operation. However, sometimes we might need another sharing key if we use a symmetric key algorithm. Therefore, the often use of the same sharing key might cause security weaknesses. In general, hash algorithms are used by authentication protocols through a public key encryption. [1]

In addition, authentication protocols ensure high security communication process in which the protocols prevent deriving the shared secret. Thus, anyhow a hacker succeeded to have the

plainext of the encryption schema, still unable to derive the secret. Important to be mentioned is that the use of hash functions is more efficient because the operating process of hashing algorithms is faster than the encryption algorithm. [10][3].

1.3.3 Digital Signature

This application uses the form of encryption applications that is considered to be a reveres of the public key encryption. It's considered being the other side of the coin of the handwritten signature in electronic media. The operating process of digital signature is as the following:

- The sender sends encrypted message, M with the secret key, d to the receiver.
- The sender also sends the signature $S(M)$ with the message.
- The receiver receives the encrypted message, decrypt it by using the public key of the sender .
- The receiver verifies the equality of the sent message to the one that had been received. [16]

Valuable to be mentioned is that during the digital signature process, there is a necessity to have a double storage area and bandwidth. The reason for that is that the amount of data of the signed messages is equal to the amount of data of the message's signature. Thus, in big messages, computation power is required since operators are taking huge amounts of power of large numbers in large modules.

Since our task is to reach a high level of efficiency and since verifying signature is a costly process. We recover the costs of the process by applying the signing standards in the mix with the hash functions. Doing so, is giving us the ability of reducing the costs of the process to achieve our goal of reaching efficiency [10].

1.4 Security Services:

From the open system interconnection (OSI) definition, the followings are considered to be the main security services:

- **Authentication:** assurance that communicating entity is the one claimed have both peer-entity and data origin authentication .
- **Access Control:** prevention of the unauthorized use of a resource .
- **Data Confidentiality:** protection of data from unauthorized disclosure.
- **Data Integrity:** assurance that the data received is as sent by an authorized entity [1].

1.5 Problem Statement

The possible contribution of this research is to develop a new hash algorithm that could present a data integrity hashing algorithm based on linear algebra. More particularly, the proposed algorithm generates non-invertible matrices by using the principle of linear combination of rows and vectors. We call this algorithm (DILH) Data Integrity using Linear Combination for Hash algorithm [17][11][6].

1.6 Methodology

This research depends on studying previous work as well as comparing existing models. A simulation of our developed model will be carried in order to make sure that expected goals are achieved.

The proposed model will be tested and simulated by using MATLAB tools, and then will be compared with an existing hash algorithm including MD5 and SHA-1 according to speed and security perspectives.

1.7 Motivation

The need of systems security integrity, confidentiality and availability leads to work on a new algorithm that works on a One-Way hash algorithm, One-Way means that it is hard to recover the original text from the hash value string. A One-Way hash algorithm is used to create digital signatures which in turn identifies and authenticates the sender and the distributed message digitally. One-Way hash functions have an important primitive cryptography, and it can be used to solve many problems including authentication and integrity [17].

Our proposed algorithm converts a variable string length into a fixed length binary sequence that cannot be reversed.

1.8 Objectives

The objectives of our research project are:

- Understand how every function of hash algorithm works and identify the strength and weakness for each of them.
- Study the One-Way hash algorithm properties and requirement.

- Develop a One-Way hash algorithm based on matrices.
- Prove that our algorithm satisfies the requirement of the One-Way hash algorithms.
- Compare our algorithm with MD5 and secure hash algorithm SHA-1, using MATLAB tool in order to test our work.
- Study the security of our new algorithm against brute force attacks, brute force attack is one in which all possible text of a certain length are tried until the correct one is found. This attack is guaranteed to work, that is why one usually chooses the length of the hash result in such a way that the brute force attack becomes impractical or too slow and thus less attractive.
- Study the security of our new algorithm against collision resistance and hamming distance.

1.9 Our Contributions

To achieve the main goal of this study, we develop a new One-Way hash algorithm called Data Integrity using Linear Combination for Hash Algorithm (DILH), which is also given a better efficiency and security, compared with a particular conventional hash algorithm. DILH algorithm using linear combination of matrices to find non-invertible matrix, that takes advantage about of the compact representation of a set of numbers in a matrix.

Hill Cipher is one of the most famous symmetric cryptosystem that can be used to protect information from unauthorized access. Hill cipher is a polygraph substitution cipher based on linear algebra, it's require inverse of the key matrix while decryption. In the fact that not all the matrices have an inverse and therefore they will not be eligible as key matrices in the Hill cipher scheme. This problem leads to many other sub problems such as the disability of decrypting any encrypted text[11][45].

In this research, We made use of the mentioned problem in Hill cipher, namely the non-invertible matrix problem, to design a new One-Way data integrity hash algorithm based on special kinds of non-invertible matrix. Particularly, we proposed to generate non-invertible key matrices by linear combinations of rows and columns of a matrix [17].

Chapter Five

One-Way hash algorithm has played an important role in modern cryptography. It is one of the indispensable tools in digital signature and authentication.

5.1 Conclusions:

In our research we:

- Designed and developed a new technique to generate non-invertible matrix from invertible matrix.
- Designed and developed a new One-Way hash algorithm based on matrix multiplications.
- Proposed a universally hashing algorithm, applicable for generating hash values. DILH could be used to detect errors, for example TCP checksums, downloaded files, and authenticate messages.
- Proved that our new One-Way Hash algorithm DILH satisfies the requirement of the One-Way hash algorithms.
- Proved that our LDIH is more efficient than other hash algorithms according to the simulations.
    - The efficiency of our hash algorithm is about 2.808 times slower than the efficiency of MD5 in 8KB file size when mv=128.
    - The efficiency of our hash algorithm is about 5.813 times slower than the efficiency of SHA-1in 8KB file size when mv=128.
    - This ratio is increased in a positive relationship with file size, so you can find that our proposed algorithm is efficient 9.794 times that SHA-1 when the file size is 256KB.
- Prove that our DILH algorithm has strong collision resistance than MD5 and SHA-1. Besides that, our algorithm shows efficiency in the hamming distance between the generated hash values.
- Prove that our DILH algorithm is strong against brute force attacks.
- Shows the optimal matrix size in all file sizes is 10X10. It has taken minimum time compared to other matrix sizes. Besides that, our algorithm didn't find any collision in this size, also it shows strength against attacks.

5.2 Future work:

This research proposed and developed a new One-Way hash algorithm that generates non-invertible matrix which cannot be reversed to produce the hash value. We proved the four requirements which the DILH algorithm needs. In the future work we will:

1. Develop our algorithm to fit in distributed systems.
2. Develop the DILH algorithm to work in any matrix not only square matrix.
3. Compare our algorithm with other hash algorithms.