Establishing the Validity of Md5 and Sha-1 Hashing in Digital Forensic Practice in Light of Recent Research Demonstrating Cryptographic Weaknesses in these Algorithms

Veronica Schmitt
Special Investigating Unit
Cyber Forensic Laboratory
Bloemfontein, South Africa

Jason Jordaan
Security and Networks Research Group
Rhodes University
Grahamstown, South Africa

ABSTRACT

MD5 and SHA-1 cryptographic hash algorithms are a standard practice in digital forensics that is used in the preservation of digital evidence and ensuring the integrity of the digital evidence. Recent studies have shown that both MD5 and SHA-1 have vulnerabilities and collisions. Based on this, the use of MD5 and SHA-1 hash algorithms in the practice of digital forensics to preserve and ensure the integrity of digital evidence has been questioned in certain instances.

Using experimentation, the researcher proves the validity of using either MD5 or SHA-1 hashing algorithms to ensure the integrity of seized digital evidence, from the moment of seizure of the evidence, through to eventual presentation and use of the evidence in court; thus demonstrating that the use of hashing remains a valid forensic methodology to ensure the integrity of digital evidence.

Keywords

Digital forensics, integrity of digital evidence, hash collisions, MD5, SHA-1, manipulation of digital evidence.

1. INTRODUCTION

Digital forensics is a scientific method in which an examiner can collect preserve and examine digital evidence while maintaining the original integrity of it [1]. The science of it needs to be aligned with the legal prerequisites for the digital evidence to be admissible in court. Recent studies have shown that MD5 and SHA-1 hash algorithms are potentially vulnerable to collisions, which could impact on the admissibility of the digital evidence in a court of law.

A hash value is the result of a mathematical calculation whereby a variable length data input is mathematically processed to produce a fixed length hash value, from which it is computationally infeasible to determine any of the input data from the resultant hash value [2]. The MD5 hash algorithm produces a 128 bit hash value, and the SHA-1 hash algorithm produces a 160 bit hash value.

The researchers will demonstrate that the use of the cryptographic hash functions remain a best practice approach for digital evidence preservation, despite the concerns identified in recent studies into MD5 and SHA-1 hash collisions, and also satisfies the legal prerequisites as set out in the Electronic Communications and Transactions Act 25 of 2002 in South Africa.

This article was presented in Lex Informatica 2013_ Establishing the Validity of Md5 and Sha-1 Hashing in Digital Forensic Practice in Light of Recent Research Demonstrating Cryptographic Weaknesses in these Algorithms

2. THE USE OF CRYPTOGRAPHIC HASH FUNCTIONS WITHIN DIGITAL FORENSICS

Hash values are an invaluable part of digital forensics, in establishing and identifying and classifying digital evidence [3]. Hash values play such an important role in the authentication and preserving the integrity of digital evidence. The use of hashing in digital forensics covers three commonly used processes and functions. The researchers have focused specifically on the use of hashing to ensure evidential integrity.

2.1 Identification of Known Files

The finding known or files from established hash sets (such as those maintained as part of the National Software Reference Library of NIST), allows the discovery of potentially incriminating data (such as child pornography images or malware) or innocent data (such as operating system files) based on the hash value of the file, rather than in the traditional sense of doing keywords and manual searches [4]. This allows an examiner to rapidly determine the nature of the data file, and to determine whether or not it requires more examination and analysis, or whether or not it is an innocuous file, which does not require further attention.

2.2 Ensure Complete Forensic Images are Made

A function of using cryptographic hashes is to determine whether all evidence has been forensically acquired from the suspect media [1]. The process followed is that a cryptographic hash (typically an MD5 or SHA-1 hash) is calculated for the suspect media, and a forensic image is then made of the suspect media. A cryptographic hash of the same type used to calculate the initial hash value is then calculated for the data in the forensic image, and this result compared to the initial hash of the suspect media, and if these match, the examiner is assured that the full suspect media has been preserved within the forensic image.

2.3 Ensure the Integrity of Evidential Data to Ensure no Data has been Altered

A crucial element within digital forensics is ensuring that the digital evidence remains unaltered from the time that it has been acquired, up until it is presented in a court of law, thereby ensuring the integrity of the evidence. The use of cryptographic hashing such as MD5 and SHA-1 and the resulting hash values have played a critical part in ensuring that and changes or alteration of digital evidence can be identified [3]. The basic premise is that even if so much is one byte in a particular file or set of data is altered after a MD5 or SHA-1 hash has been calculated for it, and it is then hashed again, it would calculate a different hash value which did not match the original.

3. MD5 AND SHA-1 HASH COLLISIONS

Hash collisions describes a situation where two different data files or data sets have a hash calculation made for them, the calculated hash values are identical, even though there are clear differences in the data themselves. Due to the nature of hash calculations, they can only provide a number of calculated values, which can naturally result in two separate data inputs resulting in the same calculated hash value. The chance of two different files randomly having the same MD5 hash value is 2^128, or a 1 in 340 billion, billion, billion, billion chance. The chance of two different files randomly having the same SHA-1 hash value is 2^160, or a 1.46 trillion, trillion, trillion, trillion chance. Identical files and data sets when hashed should always result in the same hash values.

Hash collisions can thus occur naturally from different data inputs; however the chance of this happening randomly is statistically infeasible. The concern from a digital forensics perspective is when hash collisions can be engineered so that two separate and different files return the same hash values.

3.1 Engineering Hash Collisions

In 2005 Xiaoyun Wang and Hongbo Yu of Shandong University in China presented research findings documenting how they broke the cryptographic MD 5 hash function [5]. They developed an algorithm that finds two different sequences of 128 bytes each that replicates an identical MD5 hash value. The algorithm can be used to create files of arbitrary length that will produce identical MD5 hash values, but that differed only in 128 bytes somewhere in the middle of the file [5].

Similar research managed to produce similar results with regards SHA-1 hashes [6].

This research managed to create hash collisions between two different files. These collisions of the input data for the two files being purposefully structured by the researchers in such a way that the actual mathematical processes used in the hashing process had a higher than expected probability of generate the same hash values [5] [6].

In effect, the collision resulted only as a result of using very specific input blocks, and the chance of these input blocks appearing randomly is computationally infeasible [2].

4. THE ADMISSIBILITY OF DIGITAL EVIDENCE

All cases that are adjudicated on in a court of law, whether it is a criminal prosecution or civil litigation, are dependent on admissible and relevant evidence to allow the presiding officer to make a ruling. Traditionally courts have relied on physical, real and verbal evidence to reach their findings [7]. However, in our modern world, court cases are increasingly reliant on digital evidence. The nature of digital evidence however can often pose a challenge.

4.1 Defining Digital Evidence

Digital evidence is defined as information of a legal probative value that is either stored, or transmitted in a digital form [8]. Another definition of digital evidence is that it is any data stored or transmitted using a computer that supports or refutes a theory of how an offence occurred, or addresses a critical element thereof such as intention or an alibi [8]. Digital evidence is any digital object which contains reliable information which supports or refutes a hypothesis [9]. Digital evidence includes any computer hardware (containing data), software, or data, that can be used to prove either who, what, when, where, why, and how, of an allegation being investigated [10].

4.2 The Admissibility of Digital Evidence in South African Law

The Electronic Communications and Transactions Act 25 of 2002 address the issue of digital evidence in South African law, and have allowed the use of digital evidence as evidence in a South African court of law [11]. When assigning evidential weight to digital evidence, Section 15(2) of the Electronic Communications and Transactions Act 25 of 2002 guide a court in how to evaluate the evidence [11]. A key factor to be considered in this is the reliability of the digital evidence and how the integrity of it was maintained.

In terms of current South Africa law, digital evidence, and the concept of a data message as defined in terms of the Electronic Communications and Transactions Act 25 of 2002 are synonymous. Section 1 of the Electronic Communications and Transactions Act 25 of 2002 defines data as an electronic representation of information in any form, and a data message as any data that is generated, sent, received, or stored in electronic means [12]. The Electronic Communications and Transactions Act 25 of 2002 do not define “electronic”.

Section 15 of the Electronic Communications and Transactions Act 25 of 2002 governs the admissibility and weight of data messages, and subsequently digital evidence [12]. Section 15(1) of the Electronic Communications and Transactions Act 25 of 2002 states that a data message (and thus digital evidence) cannot be ruled inadmissible simply by virtue of the evidence being in an intangible digital format, while Section 15(2) goes on to state that information in a digital form must be given due evidential weight [12].

4.3 The Role of MD5 and SHA-1 Hashing in Ensuring the Admissibility of Digital Evidence

Evidence is either admissible or inadmissible [13]. Admissible evidence is evidence that meets all regulatory and statutory requirements, and has been correctly obtained and handled [10]. The two quickest methods to ensure that evidence will not be admissible in court would be to collect it in an illegal manner, or to modify the evidence after it has come into the possession of the investigator/examiner [10].

Of significant importance are Section 15(3) of the Electronic Communications and Transactions Act 25 of 2002, which sets out guidelines for a South African court to apply in assessing the evidential weight of digital evidence [12]. This section requires a court to give due regard to:

- The reliability of the manner in which the data message (digital evidence), was generated, stored, or communicated.
- The reliability of the manner in which the integrity of the data message (digital evidence) was maintained.
- The manner in which the originator of the data message (digital evidence) was established.
- Any other relevant factor.

These factors address at a fundamental level the need for establishing a proper chain of evidence and establishing the reliability of the digital evidence using cryptographic means such as mathematical hashes [11].

In essence, the use of MD5 or SHA-1 hashes of digital evidence are used as a method to demonstrate that the evidence that is presented before court is the same as that obtained initially during the investigation, and that it has not be altered or modified in any way; thus demonstrating the integrity of the evidence.

4.4 The Potential Impact of MD5 and SHA-1 Hash Collisions on the Admissibility of Digital Evidence

The fact that MD5 and SHA-1 hashing has been potentially compromised in that files can be modified so that they produce the same hash value, raises the possibility that legal practitioners in court may argue that it does not provide adequate proof that digital evidence has not been altered from the time it has been obtained. In effect, they could argue that the digital evidence had been altered, shifting the onus onto the producing party that it had not.

5. THE VALIDITY OF MD5 AND SHA-1 HASHING IN ENSURING THE ADMISSIBILITY OF DIGITAL EVIDENCE

Mathematically, the MD5 (and SHA-1) hash calculations are of such a nature that changing one bit in any item of digital evidence will cause a cascade effect during the calculation process which would produce a different hash value [2]. The researchers conducted experimentation to validate this effect.

5.1 Experiment Methodology

The researchers' randomly selected 6175 data files from amongst the corpus of digital evidence files that were contained in all forensic images made of various digital storage media by a South African law enforcement agency digital forensic laboratory. The total number of data files which formed part of the population exceeded 100 million files.

The sample of 6175 files provided a representative sample size with a margin of error of between 1 and 2 percent [14].

Each of these files was hashed using MD5 and SHA-1, generating two separate hash values per file. These values were then documented.

Once the MD5 and SHA-1 hash values had been calculated for each file, each file was then modified making use of a hex editor to modify the files are a hexadecimal data level. For each file, the first byte of each file at logical offset 0x00 for the file was recorded, at then this byte was edited to read 0x23 or the ASCII symbol #. The file was then hashed using MD5 and SHA-1, generating two separate hash values per file. These values were then documented.

The files were then each modified again, restoring the first byte at logical offset 0x00 in the file back to its original byte value. The file was then hashed using MD5 and SHA-1, generating two separate hash values per file. These values were then documented.

5.2 Experiment Findings

For each one of the 6175 files that formed part of the sample, when the byte at logical offset 0x00 was changed, the resulting MD5 and SHA-1 hash values differed significantly from the original hash values that had been calculated. This clearly established that by changing so much as one byte within a file that a cascade effect took place, which fundamentally altered the resulting hash calculation result, thereby confirming the mathematical calculations within the respective hash functions.

When the modified byte at logical offset 0x00 in each file was returned to its original value and rehashed, the hashes that were calculated matched those made before the file had been modified. This established the reliability of the MD5 and SHA-1 hash calculations, namely that given an identical data input file into the calculation, will return an identical hash value.

6. CONCLUSION

Hash collisions can occur naturally, simply due to the number of values that can be calculated by either MD5 or SHA-1. However the chance of this occurring randomly is improbable due to the significantly large numbers involved.

While it is possible to manipulate input data in such a way that it produces two identical hash values for different inputs, the alterations have to be very specific. In other words to take an evidential file containing one set of information that proved or disproved an element of a matter before court, which had a specific hash value, and then manipulate it is such a way that it stated something else affecting the interpretation of the evidence, while still generating the same hash value, is computationally improbable.

The research has confirmed one of the key uses of both MD5 and SHA-1 as a means of demonstrating the evidential integrity of a specific digital evidence file, by showing that altering even so much as one byte of the file, that a cascade calculation process results in a significantly different hash value, making it easy to determine whether or not any data within the evidence has been altered in any way.

Based on the results of the research, it can be clearly stated that the use of MD5 and SHA-1 hashing within the field of digital forensics remains a valid scientific practice.

Legally this means that if an item of digital evidence was hashed using either MD5 or SHA-1 when it was obtained, and then hashed again using the same algorithm at a later time, and the hash values generated match, then the evidence has not been altered in the intervening time period. In other words, it the hash values match, then the integrity of the evidence from the time of acquisition to the time of presentation in court, can be relied upon.

7. REFERENCES

[1] Phillip, A., Cowen, D., & Davis, C. (2010). *Hacking Exposed Computer Forensics Second Edition.* New York: McGraw Hill.

[2] Thompson, E. (2005). MD5 Collisions and the Impact on Computer Forensics. *Digital Investigation* (2), 36-40.

[3] Prosise, C., & Mandia, K. (2003). *Incident Response and Computer Forensics* (2nd Edition). New York: McGraw Hill.

[4] Roussev, V. (2011). An Evaluation of Forensic Similarity Hashes. *The Proceedings of the Eleventh Annual DFRWS Conference* (pp. S34-S41). Elsevier.

[5] Wang, X., & Yu, H. (2005). How to Break MD5 and other Hash Functions. *Advances in Cryptology - EUROCRYPT 2005* (pp. 19-35). Berlin: Springer.

[6] Wang, X., Yin, Y. L., & Yu, H. (2005). Findings Collisions in the Full SHA-1. *CRYPTO '05 Proceedings of the 25th Annual International Conference on Advances in Cryptography* (pp. 17-36). Berlin: Springer.

[7] Joubert, C. (2001). *Applied Law for Police Officials* (2nd Edition). Lansdowne: Juta.

[8] Casey, E. (2004). *Digital Evidence and Computer Crime* (2nd Edition). London: Academic Press.

[9] Carrier, B. (2005). *File System Forensic Analysis.* Upper Saddle River: Addison-Wesley.

[10] Solomon, M. G., Barrett, D., & Broom, N. (2005). *Computer Forensics Jump Start.* Alameda: Sybex.

[11] Van Der Merwe, D., Roos, A., Pistorius, T., & Eiselen, S. (2008). *Information and Communications Technology Law.* Durban: LexisNexis.

[12] Republic of South Africa. (2002). The Electronic Communications and Transactions Act 25 of 2002. Pretoria: Government Printer.

[13] Schwikkard, P. J., & Van Der Merwe, S. E. (2002). *Principles of Evidence.* Cape Town: Juta.

[14] Saunders, M., Lewis, P., & Thornhill, A. (2009). *Research Methods for Business Students* (5th Edition). Essex: Prentice Hall.