International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

Avoidance of Sensitive Data Exposure in Intramail System Using MD5 Algorithm

Jissmol T Antony[1], Gopal B[2]

[1]PG Scholar, Dept. of Computer Science and Engineering, MITE Moodabidri, India

[2]Assistant Professor, Dept. of Computer Science and Engineering, MITE Moodabidri, India

Abstract: ***During the last few years the number of leaked sensitive data records in email system has increased dramatically). Among various data-leak cases human mistakes are one of the main causes of data loss in mail system. The existing technique is to avoid data-leak is monitor the content in storage and before transmission which is not proficient to provide protection to sensitive data and it is not capable to avoid the sensitive data leak in email system. The proposed method adds a double layer security to the sensitive data in intra mail system using the session password and comparison of the document with sensitive data repository. The amount of sensitive data in the document is verified by comparing the message digest of sensitive words in the document with the sensitive data repository where the sensitive keyword's message digest is already stored for the purpose of the comparison using MD5 algorithm. Once the comparison is complete, with the help of Term frequency process the system measures how frequently a sensitive term occurs in the selected document and produces the result.***

Keywords*:* Intramail system, Pair based matrix, MD5 Algorithm, Term frequency-Inverse document frequency, Session Password

1. Introduction

According to a report from Risk Based Security (RBS) [17], the number of leaked sensitive data records has increased radically during the last few years, i.e., from 412 million in 2012 to 822 million in 2013. Deliberately planned attacks, inadvertent leaks (e.g., forwarding confidential emails to unclassified email accounts), and human mistakes (e.g., assigning the wrong privilege) lead to most of the data-leak incidents [3]. Detecting and preventing data leaks requires a set of complementary solutions, which may include data-leak detection [2], [1], data confinement [6]–[18], stealthy malware detection [4], [7], and policy enforcement [11].

An intra mail system uses network technologies as a tool to facilitate communication among people or work groups to improve the data sharing capability and overall knowledge base of an organization's employees. An intra mail system uses technologies of internet protocol to share any type of information within association or Business Company. This technology is used in the proposed system, the Intra Mailing System to make the most out of this technology. This will somehow automate the information passing system.

As organization grows in size in terms of departments and functionalities, it requires a quick and efficient system to achieve instant communication between employees of same department or between departments. The Intra Mailing System serves organization"s needs in a consistent and transparent manner. It should cater the needs of information sharing. It allows the users to exchange their views through mails and send electronic files thru attachments. It should have all traditional things such as sent items, inbox, drafts etc. The users are allows to send mails to multiple users and groups too. Thus the system caters spontaneous needs of the organization.

Like the communication between employees among same department or between different department increases, there is a possibility to increase in data leaks or sensitive data exposure. This data leak may happen due to employees fault. Inadvertent data leak may cause a harmful effect to organization or to a particular department.

The method here proposed is provides additional security to the user sensitive data. As the first layer of security we included the image based session password concept where the user needs to enter the session password when the new session starts. Pair based matrix technique we used in the session password. Along with session password user has to enter his/her Id and password which is created during user registration.

Once the user log in to the system he can send the mail to the person who he want to. He/she has to select the document file. The selected document file undergoes comparing with the sensitive data repository which is maintained by Admin. After comparison finishes the result will be produced with the help of Term frequency-Inverse document frequency algorithm.

2. Session Password

A session password is a password that is valid for only one login session. The most important advantage that is addressed by session password is that, in contrast to static passwords, they are not vulnerable to replay attacks. This means that a potential intruder who manages to record session passwords that was already used to log in to a service will not be able to abuse it, since it will no longer be valid. A second major advantage is that a user, who uses the same (or similar) password for multiple systems, is not made vulnerable on all of them, if the password for one of these is gained by an attacker.

A. Pair Based Matrix

Pair-based Authentication scheme is used in session password. Each time of login user has to select few images

from displayed. Based on the images selected, the session password generated. Whenever user logins into the system every time new session is created and that session remains as it is until user gets log out.

3. MD5 algorithm

The MD5 message-digest algorithm is a widely used vulnerable cryptographic hash function producing a 128-bit (16-byte) hash value, typically expressed in text format as a 32-digit hexadecimal number. MD5 has been utilized in a wide variety of cryptographic applications and is also commonly used to verify data integrity. MD5 is a one-way function; it is neither encryption nor encoding. It can be reversed by brute-force attack and suffers from extensive vulnerabilities.

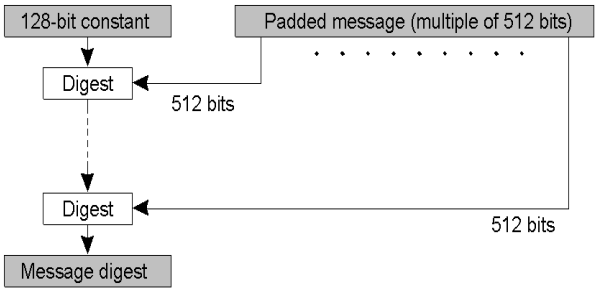

Figure 1: MD5 Algorithm Structure

A. MD5 Steps

- Step 1: Append padded bits

The message is padded so that its length is congruent to 448, modulo 512.vMeans extended to just 64 bits shy of being ofv512 bits long. A single "1" bit is appended to the message, and then "0" bits are appended so that the length in bits equals 448 modulo 512.

- Step 2: Append length

64 bit representation of b is appended to the result of the previous step. The resulting message has a length that is an exact multiple of 512 bits.

- Step 3: Initialize MD Buffer

A four-word buffer (A, B, C, and D) is used to compute the message digest. Here each of A, B, C, D, is a 32 bit register. These registers are initialized to the following values in hexadecimal:
Word A: 01 23 45 67
Word B: 89 ab cd ef
Word C: fe dc ba 98
Word D: 76 54 32 10

- Step 4: Process message in 16-word blocks.

Four auxiliary functions that take as input three 32-bit words and produce as output one 32-bit word.
F(X,Y,Z) = XY v not(X) Z
G(X,Y,Z) = XZ v Y not(Z)
H(X,Y,Z) = X xor Y xor Z
I(X,Y,Z) = Y xor (X v not(Z))
Process message in 16-word blocks cont. If the bits of X, Y, and Z are independent and unbiased, the each bit of F(X,Y,Z), G(X,Y,Z), H(X,Y,Z), and I(X,Y,Z) will be independent and unbiased.

- Step 5: Output

The message digest produced as output is A, B, C, and D. That is, output begins with the low order byte of A, and end with the high-order byte of D.

The MD5 algorithm is simple to implement, and provides a "fingerprint" or message digest of a message of arbitrary length. The difficulty of coming up with two messages with the same message digest is on the order of 2^64 operations.

4. Term Frequency Algorithm

Term frequency-inverse document frequency is a numerical statistic that is intended to reflect how important a word is to a document. The tf-idf weight is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. Variations of the tf-idf weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query.

A. Term Frequency

Term Frequency measures how frequently a term occurs in a document. Since every document is different in length, it is possible that a term would appear much more times in long documents than shorter ones. Thus, the term frequency is often divided by the document length (aka. the total number of terms in the document) as a way of normalization:
TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document).

B. Inverse Document Frequency

Inverse Document Frequency measures how important a term is. While computing TF, all terms are considered equally important. However it is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms while scale up the rare ones, by computing the following:
IDF(t) = log_e(Total number of documents / Number of documents with term t in it).

5. Conclusion

The proposed approach is a sensitive data-leak detection model and presents its realization. Using the technique, the exposure of the sensitive data is kept to a minimum during the detection. We have conducted extensive experiments to validate the accuracy, privacy, and efficiency of our solutions. For future work, we plan to focus on designing a host-assisted mechanism for the complete data-leak detection for large-scale organizations.

6. Acknowledgment

We acknowledge all our gratitude to Prof. Dr. Nagesh H.R (Dean and Head of Computer Science and Engineering Department) who has provided facilities to explore the subject with more enthusiasm. This experience will always steer us to do our work perfectly and professionally.

References

[1] K. Borders and A. Prakash, "Quantifying information leaks in outbound web traffic," in Proc. 30th IEEE Symp. Secur. Privacy, May 2009, pp. 129–140.
[2] Identity Finder. Discover Sensitive Data Prevent Breaches DLP Data Loss Prevention. [Online]. Available: http://www.identityfinder.com/, accessed Oct. 2014.
[3] X. Shu and D. Yao, "Data leak detection as a service," in Proc. 8th Int. Conf. Secur. Privacy Commun. Netw., 2012, pp. 222–240.
[4] A. Kapravelos, Y. Shoshitaishvili, M. Cova, C. Kruegel, and G. Vigna, "Revolver: An automated approach to the detection of evasiveweb-based malware," in *Proc. 22nd USENIX Secur. Symp.*, 2013, pp. 637–652.
[5] A. Nadkarni and W. Enck, "Preventing accidental data disclosure in modern operating systems," in Proc. 20th ACM Conf. Comput. Commun. Secur., 2013, pp. 1029–1042.
[6] A. Kapravelos, Y. Shoshitaishvili, M. Cova, C. Kruegel, and G. Vigna,"Revolver: An automated approach to the detection of evasiveweb-based malware," in Proc. 22nd USENIX Secur. Symp., 2013, pp. 637–652.
[7] X. Jiang, X. Wang, and D. Xu, "Stealthy malware detection and monitoring through VMM-based „out-of-the-box" semantic view reconstruction," ACM Trans. Inf. Syst. Secur., vol. 13, no. 2, 2010, p. 12.
[8] Y. Jang, S. P. Chung, B. D. Payne, and W. Lee, "Gyrus: A framework for user-intent monitoring of text-based networked applications," in Proc. 23rd USENIX Secur. Symp., 2014, pp. 79–93.
[9] S. Geravand and M. Ahmadi, "Bloom filter applications in network security: A state-of-the-art survey," Comput. Netw., vol. 57, no. 18, pp. 4047–4064, Dec. 2013.
[10] B. Wang, S. Yu, W. Lou, and Y. T. Hou, "Privacy-preserving multikeyword fuzzy search over encrypted data in the cloud," in Proc. 33th IEEE Conf. Comput. Commun., Apr./May 2014, pp. 2112–2120.
[11] G. Karjoth and M. Schunter, "A privacy policy model for enterprises," in *Proc. 15th IEEE Comput. Secur. Found. Workshop*, Jun. 2002, pp. 271–281.malware," in *Proc. 22nd USENIX Secur. Symp.*, 2013, pp. 637–652.
[12] J. Croft and M. Caesar, "Towards practical avoidance of information leakage in enterprise networks," in Proc. 6th USENIX Conf. Hot Topics Secur. (HotSec), 2011, p. 7.
[13] M. Burkhart, M. Strasser, D. Many, and X. Dimitropoulos, "SEPIA: Privacy-preserving aggregation of multi-domain network events and statistics," in Proc. 19th USENIX Conf. Secur. Symp., 2010, p. 15.
[14] X. Yi, M. G. Kaosar, R. Paulet, and E. Bertino, "Single-database private information retrieval from fully homomorphic encryption," IEEE Trans. Knowl. Data Eng., vol. 25, no. 5, pp. 1125–1134, May 2013.
[15] G. Jagannathan and R. N. Wright, "Privacy-preserving distributed k-means clustering over arbitrarily partitioned data," in Proc. 11th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2005, pp. 593–599.
[16] J. Jung, A. Sheth, B. Greenstein, D. Wetherall, G. Maganis, and T. Kohno, "Privacy oracle: A system for finding application leaks with black box differential testing," in Proc. 15th ACM Conf. Comput. Commun. Secur., 2008, pp. 279–288.
[17] GXiaokui Shu, Danfeng Yaoand Elisa Bertino, "Privacy-Preserving Detection of Sensitive Data Exposure," IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, VOL. 10, NO. 5, MAY 2015
[18] H. Yin, D. Song, M. Egele, C. Kruegel, and E. Kirda, "Panorama: Capturing system-wide information flow for malware detection and analysis," in *Proc. 14th ACM Conf. Comput. Commun. Secur.*, 2007, pp. 116–127.