IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 6, June 2016

Study on Privacy-Preserving and Detection of Sensitive Data Exposure using Message Digest

Ashwini T K[1], Mohan Kumar S[2], Jagadeesh S N[3]

Department of Information Science and Engineering, MSRIT, Bangalore, India

Abstract: Analysis from security bodies, research institutes, government organizations shows that the amount of data leak incidents is more now a day. Most of the data leak incidents is because of mistakes done by humans, such type of leaks can be classified as inadvertent data leak.In this paper we present a privacy preserving data-leak detection (DLD) method to solve the data leak issue such as human mistakes (data loss through e-mail). We presented the design method using fuzzy fingerprint (Message Digest5 algorithm) technique that enhances data privacy during data leak detection operation. This approach is based on fast and real-world one-way encryption on the sensitive data. The approach enables the data owner to safely delegate the detection operation to the DLD provider by generating set of message digests to reduce inadvertent data leak in network traffic. This method assists the organization or owner to know the details of the file whether the file is sensitive or not through displaying or sending an alert message.

Keywords: Data leak detection (DLD), Message Digest (MD5), Social security number (SSN), Electronic mail (E-Mail).

I. INTRODUCTION

The government organizations and research institutions shows that the information or data leaks are growing rapidly in recent years. Data leakage is defined as the accidental or unintentional distribution of private or sensitive data to an unauthorized entity. Sensitive data can be defined as the information that relates as a consumer, client, employee, patient and student. It can be identifying information as well our contact information, identification, numbers and sensitive data such as date of birth, expiry date etc.

The sensitive data in companies and organizations includes intellectual property, financial information, personal credit card data, and other sensitive data depending on the business. Data leakage poses a serious issue for companies, organizations and research institutions. It is enhanced by the fact that transmitted data such as emails, instant messaging, website forms and file transfers are largely regulated and unmonitored on their way to their destinations.

Furthermore in many cases, sensitive data are shared among various stakeholders such as employees working from outside the organization’s premises, business partners and customers. This increases the risk that confidential information will fall into unauthorized hands [1].

Data leak can be grouped into inadvertent data leak, malicious data leak, legitimate & intended data transfer. The inadvertent (human mistakes) data leaks are one of the main reasons of data leakage. In order to detect a data leakage of inadvertent sensitive data leaks caused by human mistakes a common approach is to screen content in storage and transmission for exposed sensitive information such an approach normally requires the detection operation to be conducted in secrecy. This secrecy is challenging to satisfy in practice, as detection servers may be compromised or outsourced. The proposed method is that it enables the data owner to get an alert if the file is sensitive before forwarding it to the intended recipient [1] [3].

Over the last few years, companies in every industry sector around the globe have found their sensitive internal information lost or leaked to the outside world. According to a report from Risk BasedSecurity (RBS), the number of leaked sensitivedata records has increased dramatically during the lastfew years thatare from 412 million in 2012 to 822 millionin 2013.

Among them deliberately planned attacks, inadvertent leaks such as forwarding confidential emails to unclassified emailaccounts and human mistakes are assigning the wrongprivilege lead to most of the data-leak incidents. As data is likely one of organizations most valuable assets, protecting it out of the public domain plays are major roles. In order to achieve data leakage a different methods of protections are taken place in the literature [1] [10].

Electronic mail data transfer is very much essential for an organization’s day to day operationsbut it becomeleads to main sources of data leakage according to a Ponemon institute report. In a survey report of 830 information technology, security and compliance professionals, more than half of the respondents said, improper email use by employees are main cause of data leakage.

The survey report found that approximately 69 percent employees are reported that have violated security policies and frequently send sensitive information through insecure email channels, and 60 percent use personal webmail accounts to send corporate information. About 63 percent believe employees mistakenly send confidential

DOI 10.17148/IJARCCE.2016.5624

information to recipients outside the workplace. In addition, 70 percent of the compliance and security professionals surveyed are concerned about data loss via email on mobile devices [11].

Data leak prevention (DLP) is a suite of technologies aimed at stemming the loss of sensitive information that occurs in enterprises across the globe. By focusing on the location, classification and monitoring of information at rest, in use and in motion, this solution can go far in helping an enterprise get a handle on what information it has, and in stopping the numerous leaks of information that occur each day [7][8].

Detecting and preventing data leaks requires a set ofcomplementary solutions, which may include data-leakdetection, data confinement, stealthy malwaredetection, and policy enforcement. Network data-leak detection (DLD) typically performs deep packet inspection (DPI) and searches for many occurrences of sensitive data patterns. DPI is a technique to analyse payloads of TCP/IP packets for inspecting application layer data (HTTP header/content). The alerts are triggered when the amount of sensitive data found in traffic passes a threshold. The detection system can be deployed on a router or integrated into standing network intrusion detection systems (NIDS) [1] [12] [14].

The standing solution of data leak detection requires the plaintext sensitive data. Still this solution is undesirable, as it may threaten the confidentiality of the sensitive information. If a detection system is negotiated with the more occurrences of sensitive data patterns then it may expose the plaintext sensitive data (in memory). The data owner may need to outsource the data-leak detection to the providers but he may unwilling to reveal the plaintext sensitive data to them. Therefore, one needsnew data-leak detection solutions that allow the providers to scan content for leaks without learning the sensitive data. There is less optimal solution for standing data leak detection system, it causes less confirmation, no data confinement and less security [1].

II. PREVIOUS RELATED WORK

In this section we present analysis of previous related work carried out for the privacy preserving data leakage detection problems with a threat model, a security goal and a privacy goal.

A. Data owner or organization owns the sensitive data handles the DLD provider to inspect the network traffic from the organizational networks for inadvertent data leak. However, the data owner does not want to reveal the sensitive data to the DLD provider directly.

B. DLD provider inspects the network traffic for potential data leaks. The inspection can be performed offline without causing any real-time delay in routing the packets. However, the DLD provider may attempt to gain knowledge about the sensitive data [1] [2].

C. Security Goal, Threat Model & Privacy Model

Case 1-Inadvertent data leakage:The sensitive data is accidentally leaked to the outside world by a unauthorized user. This paper focuses on type of data leaks, the main causes of inadvertent data leak includes human errors such as forgetting to use encryption, carelessly forwarding an internal mail and attachments to outsiders, or due to application flaws.

Case 2-Malicious data leakage: A piece of stealthy software may steal sensitive personal or organizational information from a host because malicious adversary can use strong encryption or steganography to disable content based traffic inspection, thus this type of leaks are out of scope of our network based solution .

Case 3-Legitimate and intended data transfer: The sensitive data is sent by a legitimate user for legitimate purposes. In this paper, we assume that legitimate data transfers use data encryption such as SSL, which allows one to distinguish it from the inadvertent data leak. Therefore, in what follows we assume that plaintext sensitive data appearing in network traffic is only due to inadvertent data leaks [1] [2] [5].

The security goal in this paper is to detect Case 1 leaks that are inadvertent data leaks. In other words, we aim to detect sensitive data appearance in traffic (attached email) over supervised network channels. We assume that 1) plaintext data in supervised network channels can be extracted for inspection 2) the data owner is aware of legitimate data transfers 3) whenever sensitive data is found over network traffic, the data owner can decide whether or not it is a data leak [2] [5].

The privacy goal in our fuzzy fingerprint mechanism is to prevent the DLD provider from inferring the exact knowledge of the sensitive data; the DLD provider is given the fingerprints of sensitive data and the content of network traffic which may or may not contain data leak. In our model, we aim to hide the sensitive values among other no sensitive values, so that the DLD provider is unable to pinpoint sensitive data among them even under data-leak scenarios [1] [2].

Our privacy goal is defined as follows. The DLD provider is given digests of sensitive data from the data owner and the content of network traffic to be examined. The DLD provider should not find out the exact value of a piece of sensitive data with more than 1/K probability, where K is aninteger representing the number of all possible sensitive-data candidates that can be inferred by the DLD provider [1] [2] [5].

III.SYSTEM DESIGN

In this section we describe the proposed fuzzy based fingerprint mechanisms for privacy preserving data leak detection method. The proposed system architecture consists of two phases respectively that is shingles, fingerprints and fingerprint filter extensions.

Shingles & Fingerprint: Noise tolerance is realized by through the use of shingles. It refers to the fixed size sequence of contiguous characters. For example, for string abcdefgh, the 3-gram shingle set consists of six elements {abc, bcd, cde, def, efg, fgh}. A sliding window is used in

DOI 10.17148/IJARCCE.2016.5624

shingling a document for finding duplicate web documents. One needs to transform each shingle element into its digests or fingerprints which uniquely represents the data.

For this purpose we use MD5, it is one of the most widely used cryptographic hash functions now a day. It can compress any length of data into an information digest of 128 bits while this segment message digest often claims to be a digital fingerprint of the data. This algorithm makes use of a series of non-linear algorithm to do the circular operation, so that cracker cannot restore the original data. It can effectively prevent data leakage caused by inverse operation.

Fingerprint Filter Extensions:

The amount of data all over the world is increasing day by day, hence we secure these type of data by generating fingerprints or message digest it is called hashing technique we implemented fingerprints and extensions. This extension to use Bloom filter in the DETECT operation for efficient set intersection set. Bloom filter is a well know space saving data structure for performing set-membership test. It applies multiple hash functions to each of the set elements and stores the resulting values in a bit vector. Bloom filter is combination with Rabin fingerprint is referred to us as the fingerprint filter. We consider Rabin fingerprints with variety of module's in fingerprint filter as the hash objective function implementation (MD5) and we perform extensive experimental setup and evaluation on both the approach with MD5 which is presented in section V.

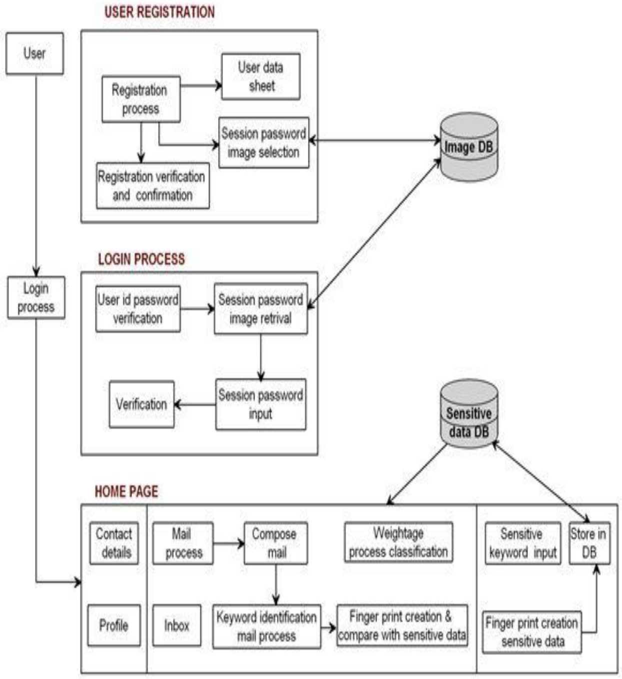
USER REGISTRATION
User
Registration process
User data sheet
Session password image selection
Registration verification and confirmation
Image DB
LOGIN PROCESS
Login process
User id password verification
Session password image retrival
Verification
Session password input
Sensitive data DB
HOME PAGE
Contact details
Mail process
Compose mail
Weightage process classification
Sensitive keyword input
Store in DB
Profile
Inbox
Keyword identification mail process
Finger print creation & compare with sensitive data
Finger print creation sensitive data

Figure 1: The proposed system architecture

Figure1 describe the proposed model which consists of three modules that is User Registration, Login Process, Homepage modules. The modules performs the below activities, Creating Session password using session ids during User Registration-in registration module user has a predefined image used as input for password generation which generates row and column value used for session password creation. This session password creates every time user logs, a random function deployed as an objective for the module to make secure session for the user.

Adding sensitive keywords and weightage-in this module admin adds sensitive keywords with its weightage, these sensitive keywords processed into fingerprints hash code by using MD5 algorithm approach for data privacy.

Detecting Sensitive Data in the text file-In this module a ASCII text file is used which contents sensitive data. Through FTP or uploading to remote server or sending, a special characters removal process and generation of all keywords into fingerprints hash code process takes place, if it matches with sensitive word fingerprints then take the weightage of those sensitive words and compare with the threshold value. If the sensitive data weightage is greater than the threshold value, DLD provider will give alert message to owner or organization.

IV. ANALYSIS AND DISCUSSION

In this section we present the analysis of security and privacy guarantees provided by considering in the section 2 data-leak detection systems. In the following section we present privacy analysis. We consider static web server in our implementation in order to extract the sensitive data and match the digest in the considerable Poisson distribution network traffic. We identify the limitations associated with proposed approach.

A. Privacy Analysis

Our privacy goal is to prevent the DLD provider from inferring the exact knowledge of all sensitive data, both the outsourced sensitive data and the matched digests in network traffic. We quantify the probability for the DLD provider to infer the sensitive shingles as follows.

A polynomial-time adversary has no greater than $\frac{2^{(\mathrm{pf}-\mathrm{pd})}}{\mathrm{n}}$ probability of correctly inferring a sensitive shingle, where pf is the length of a fingerprint in bits, pd is the fuzzy length, and n $\in 2^{(pf-pd)},\left(2^{pf}\right)]$ is the size of the set of traffic fingerprints, assuming that the fingerprints of shingles are uniformly distributed and are equally likely to be sensitive andappear in the traffic. Where K is the fingerprints matched for fuzzy fingerprints.

$$\mathrm{k}=\frac{n}{2^{pf}} * 2^{pd} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

In the next section we determine and present expected alert rate by calculating sensitive word weighted average method for sensitive data analysis as follows.

Alert Rate: We evaluate the alert rate for each sensitive word by calculating the sensitive word weightage, $\boldsymbol{S}_{\boldsymbol{wc}}$ that is,

$$\boldsymbol{S}_{\boldsymbol{wc}} = \frac{\boldsymbol{N}_{\boldsymbol{osw}}}{\boldsymbol{T}_{\boldsymbol{nf}} * \boldsymbol{K}_{\boldsymbol{w}}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots. (2)$$

DOI 10.17148/IJARCCE.2016.5624

Where N_{osw}Number of occurrences of sensitive word, T_{nf} total number of keywords in a file and K_w key weightage.
The expected alert rate R is expressed as

$$R= (S_1+S_2+S_3\ldots+S_n) > S_{th}\ldots\ldots\ldots\ldots(3)$$

Where S_1, S_2, $S_3\ldots$, S_n are individual sensitive words weightage and S_{th} is Threshold. If sum of the sensitive words weightage is greater than S_{th} then organization gets an alert message.
Performance analysis of hashing algorithms
The cryptography is a way of securing the information and message over the internet. The irreversible form of hashing technique is MD5 which is used to measure the data integrity using 128 bit message, it is then handover to user to generate a fingerprint message. Hence MD5 algorithm is the optimal solution for generation of message than SHA algorithm.

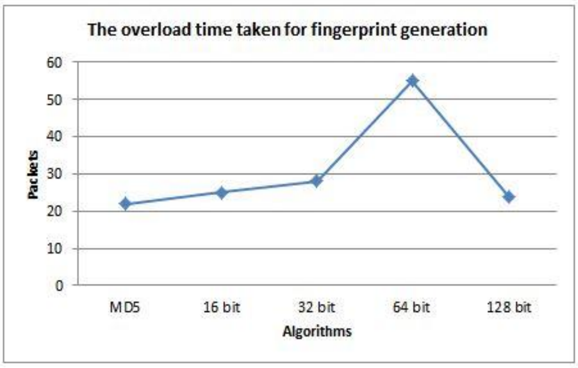

Figure 2: The overload time taken for fingerprint generation

We plotted the overload time taken for fingerprint generation using MD5 16bit, 32bit, 64bit and 128bit versus packet as shown in the figure 2. It is seen that 5 different digest size has been consider as an input data to generate hash key functions for matching fuzzy fingerprints with a different intervals of performance key functions such as 2, 6, 10 hash function. We inspect 20 to 60 packets with approximately 400 words comparatively. From the above graph it is clear that message digest5 is optimally a good approach for fingerprints generation.

V. EXPERIMENTAL EVALUATION AND DISCUSSION

We setup three experimental setup in order to find the fingerprints matched percentage.
Exp.1 True leak-a user leaks the entire set of sensitive data via FTP by uploading it to a remote FTP server.
Exp.2 No leak-the non-related outbound HTTP traffic of 20 users is captured (30 minutes per user), and given to the DLD server to analyse. No sensitive data (i.e., zero true positive) should be confirmed.
Exp.3 No leak-The Enron dataset (2.6 GB data, 150 users' 517,424 emails) as a virtual network traffic is given to the DLD server to analyse [1][2][5].

We implemented entire application and tested experimental setup with intranet facility. We used Java as a front end module and MySql as a backend module. The resultant segment message digest claims optimally better compare to SHA algorithm from the figure for digital fingerprints of data [3].

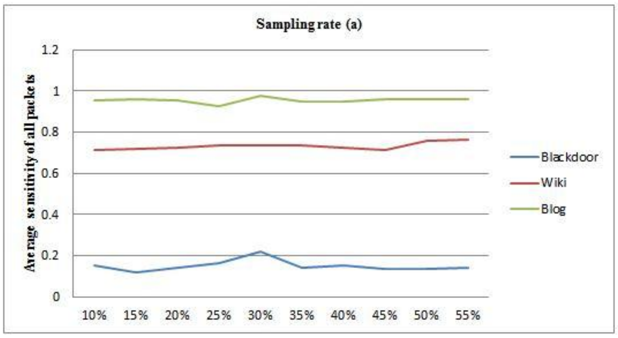

Figure 3: Detection accuracy comparison in terms of average sensitive packets

In figure3 the sensitivities of tests differ due to dissimilar levels of changes by the leaking programs, which make it hard to perform detection. Word Press alternates spaces with +'s when directing the HTTP POST request [1].

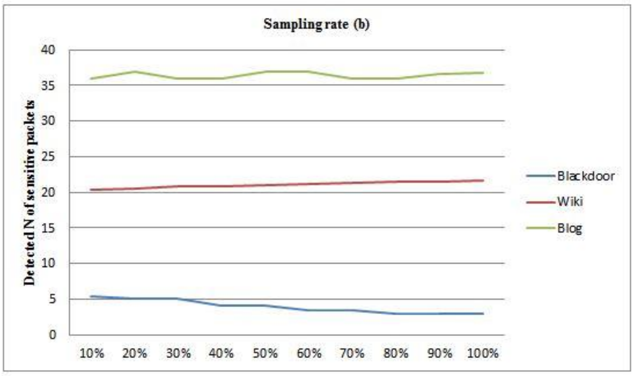

Figure 4: Detection accuracy comparison in terms of number of sensitive packets

In figure4 the outcomes cover both outbound and inbound traffic flow and dual the real number of sensitive packets in Blog and Wiki setups due to HTML fetching and presenting of the submitted data.

The figure 3 and 4 shows detection accuracy evaluation in terms of Sampling rate (a) the averaged sensitivity and Sampling rate (b) the amount of perceived sensitive packets. X-axis is the part of revelation rate that is the part of sensitive-data fingerprints exposed to the DLD server.

The outcomes indicate that the use of limited sensitive-data fingerprints does not much reduce the detection ratio related to the use of complete sets of sensitive-data fingerprints. On the other hand, extreme lesser sampling rates example 10%, may not deliver adequate amounts of fingerprints to define the leaking characteristic of the traffic flow.

DOI 10.17148/IJARCCE.2016.5624

VI. CONCLUSION AND FUTURE WORK

We proposed fuzzy fingerprint (MD5), a privacy-preserving data leak detection method using message digest technique and present its realization. Using special digests, the exposure of the sensitive data is kept to a minimum during the detection. We have set up and conducted experiments to validate the simple leaking scenario like true leakage by remote server access, outbound HTTP traffic and privacy of proposed method. Future focus on designing a host assisted mechanism, one-way encryption approach for the remote server and cloud based server for the complete data-leak detection for large scale organizations.

ACKNOWLEDGMENT

As a result of this paper were decided our future and roadmap of post-graduation studies at M S Ramaiah Institute of Technology, affiliated to VTU Belgaum, Karnataka, India. At research centre M S Ramaiah Institute of Technology, Bangalore, Karnataka, India.

REFERENCES

[1] Shu, Xiaokui, Danfeng Yao, and Elisa Bertino. "Privacy-preserving detection of sensitive data exposure." Information Forensics and Security, IEEE Transactions on 10.5 (2015): 1092-1103.
[2] Shu, Xiaokui, and Danfeng Daphne Yao. "Data leak detection as a service."Security and Privacy in Communication Networks. Springer Berlin Heidelberg, 2012. 222-240.
[3] Yong-Xia, Zhao, and Zhen Ge. "MD5 research." 2010 Second International Conference on Multimedia and Information Technology. IEEE, 2010.
[4] Wang, Xiaoyun, and Hongbo Yu. "How to break MD5 and other hash functions." Advances in Cryptology–EUROCRYPT 2005. Springer Berlin Heidelberg, 2005. 19-35.
[5] Shu, Xiaokui, and Danfeng Daphne Yao. "Data leak detection as a service: challenges and solutions." Virginia Tech2012 (2012).
[6] Gupta, Piyush, and Sandeep Kumar. "A Comparative Analysis of SHA and MD5 Algorithm." architecture 1 (2014): 5.
[7] Liu, Simon, and Rick Kuhn. "Data loss prevention." IT professional 12.2 (2010): 10-13.
[8] Karjoth, Günter, and Matthias Schunter. "A privacy policy model for enterprises." Computer Security Foundations Workshop, 2002. Proceedings. 15th IEEE. IEEE, 2002.
[9] Broder, Andrei Z. "Some applications of Rabin’s fingerprinting method."Sequences II. Springer New York, 1993. 143-152.
[10] Risk Based Security. (Feb. 2014). Data Breach Quick- View: An Executive’s Guide to 2013 Data Breach Trends. [Online]. Available: https://www.riskbasedsecurity.com/reports/2013-DataBreachQuickView.pdf, accessed Oct. 2014.
[11] Ponemon Institute. (May 2013). 2013 Cost of Data Breach Study: GlobalAnalysis. [Online].Available:https://www4.symantec.com/mktginfo/whitepaper/053013_GL_NA_WP_Ponemon-2013-Cost-of-a-Data-Breach-Report_daiNA_cta72382.pdf, accessed Oct. 2014.
[12] Identity Finder. Discover Sensitive Data Prevent Breaches DLP DataLoss Prevention. [Online]. Available: http://www.identityfinder.com/, accessed Oct. 2014.
[13] G. Karjoth and M. Schunter, “A privacy policy model for enterprises,”in Proc. 15th IEEE Comput. Secur. Found. Workshop, Jun. 2002,pp. 271–281.
[14] Symantec. Data Loss Prevention (DLP) Software. [Online]. Available:http://www.symantec.com/data-loss-prevention/, accessed Oct. 2014.
[15] Global Velocity Inc. Cloud Data Security From the Inside Out. [Online]. Available: http://www.globalvelocity.com/, accessed Oct. 2014.
[16] A. Z. Broder, “Some applications of Rabin’s fingerprinting method,” in Sequences II. New York, NY, USA: Springer-Verlag, 1993, pp. 143–152.
[17] SN, Jagadeesha et al, “Routing in all optical networks using recursive state space technique”, signal & image processing An International Journal vol 7, Number 2, April 2016, pp 23-37.
[18] Mohan Kumar S and Jagadeesh S N, “Routing and Wavelength Assignment for Networks using Link State Space Technique”, in Proceedings of the 6th IETE National Conference on RF & Wireless, 09-11 May 2013 Shimoga Karnataka pp 116-119.
[19] Mohan Kumar S, Jagadeesh S N and Swaroop “ A case study on Markov model for double fault tolerance, comparing cloud based storage system” IJARCCE vol 4, Issue 11, November 2015, DOI 10.17148 pp 240-245.

BIOGRAPHIES

Ashwini T K received Bachelor of Engineering in Computer Science and Engineering in 2014 from VTU Belgaum, Karnataka, India. She is pursuing Master Degree (Software Engineering) at M S Ramaiah Institute of Technology, Bangalore. Her interest area is Web Security.

Mohan Kumar S is currently working as Assistant Professor, Department of Information Science and Engineering, M S Ramaiah Institute of Technology, Bangalore, Karnataka, India.

DOI 10.17148/IJARCCE.2016.5624