FAST DETECTION OF TRANSFORMED DATA LEAKS ON MAIL SERVER USING LEVENSTEIN-DISTANCE TECHNIQUES

K.Jaisharma[1] | S.Yogeswari[2] | T.Deepika[3]
[1](Department of CSE, Assistant Professor, GKM College of Engineering and Technology, Perungalathur, Chennai-63)
[2](Department of CSE, GKM College of Engineering and Technology, Perungalathur, Chennai-63)
[3](Department of CSE, GKM College of Engineering and Technology, Perungalathur, Chennai-63)

Abstract— *The main aim of this project is to provide security against sensitive data, using data leak detection technique on the transformed data. The leak of sensitive data on computer systems poses a serious threat to organizational security. Statistics show that the lack of proper encryption on files and communications due to human errors is one of the leading causes of data loss. Organizations need tools to identify the exposure of sensitive data by screening the content in storage and transmission,i.e., to detect sensitive information being stored or transmitted in the clear. However, detecting the exposure of sensitive information is challenging due to data transformation in the content. Transformations (such as insertion and deletion) result in highly unpredictable leak patterns. In this paper, we utilize sequence alignment techniques for detecting complex data-leak patterns. Our algorithm is designed for detecting long and inexact sensitive data patterns. This detection is paired with a comparable sampling algorithm, which allows one to compare the similarity of two separately sampled sequences. Our system achieves good detection accuracy in recognizing transformed leaks. We implement a parallelized version of our algorithms in graphics processing unit that achieves high analysis throughput. We demonstrate the high multithreading scalability of our data leak detection method required by a sizable organization.*

Keywords—Data Leak Detection;Parallelism;Alignment;Sampling;Dynamic Programming

1. INTRODUCTION

Statistics from security firms, research institutions and government organizations show that the number of data-leak instances has grown rapidly in recent years. Among various data-leak cases, human mistakes are one of the main causes of data loss. According to a report from Risk Based Security (RBS) the number of leaked sensitive data records has increased dramatically during the last few years. Deliberately planned attacks, inadvertent leaks (forwarding confidential emails to unclassified email accounts) and human mistakes (assigning the wrong privilege) lead to most of the data-leak incidents. Detecting and preventing data leaks requires a set of complementary solutions, which may include data-leak detection data confinement stealthy malware detection and policy enforcement. Network data-leak detection (DLD) typically performs deep packet inspection (DPI) and searches for any occurrences of sensitive data patterns.

In our data leak detection model, we analyze two types of sequences: sensitive data sequence and content sequence.

- Content sequence is the sequence to be examined for leaks. The content may be data extracted from file systems on personal computers, workstations, and servers; or payloads extracted from supervised network channels(details are discussed below).
- Sensitive data sequence contains the information (e.g.,customers' records, proprietary documents) that needs
- to be protected and cannot be exposed to unauthorized parties. The sensitive data sequences are known to the
- analysis system.In this paper, we focus on detecting inadvertent data leaks,and we assume the content in file system or network traffic(over supervised network channels) is available to the inspection system. A supervised network channel could be an unencrypted channel or an encrypted channel where the content in it can be extracted and checked by an authority. Such a channel is widely used for advanced NIDS where MITM

(man-in-the-middle) SSL sessions are established instead of normal SSL sessions. We do not aim at detecting stealthy data leaks that an attacker encrypts the sensitive data secretly before leaking it. Preventing intentional or malicious data leak,especially encrypted leaks, requires different approaches and remains an active research problem.

2. EXISTING SYSTEM:

In an existing system straight forward realizations of data-leak detection require the plaintext sensitive data. However, this requirement is undesirable, as it may threaten the confidentiality of the sensitive information. If a detection system is compromised, then it may expose the plaintext sensitive data. In addition, the data owner may need to outsource the data-leak detection to providers, but may be unwilling to reveal the plaintext. There was no privacy preserving in existing system, so providers can access the data without data-owner's permission.

3. PROBLEM DEFINITION:

1. *Data traffic on proxy server and mail server Impacts the performance of data Inadvertent data leak :*

The sensitive data is accidentally leaked in the outbound traffic by a legitimate user. Inadvertent data leak may be due to human errors such as forgetting to use encryption, carelessly forwarding an internal email and attachments to outsiders.

2. *Malicious data leak :*

A rogue insider or a piece of stealthy software may steal sensitive personal or organizational data from a host.

3. *Data Traffic and Time consumption:*

Leakage detection technique and time delay of common legitimate users.

4. *Static filtering of authorized users:*

Static approaches of authorized user filtering technique affect the efficient of data leakage detection.

4. PROPOSED SYSTEM:

In our proposed system we propose a data-leak detection solution which can be outsourced from organization, we design and implement Lucene search engine framework Levenshtein-distance technique to avoid data leak and also provide privacy preserving to sensitive data. Two most important players in our proposed model is

- Data Owner owns the sensitive data and authorizes the DLD provider to inspect the network traffic from the organizational networks for anomalies, namely inadvertent data leak.
- Mail Server - DLD provider inspects the network traffic for potential data leaks. We focus on detecting inadvertent data leaks, and we assume the content in file system or network traffic (over supervised network channels) is available to the inspection system. A supervised network channel could be an unencrypted channel or an encrypted channel where the content in it can be extracted and checked by an authority. Authority has the threshold for every categorized position of users.
- In our security model, we assume that the analysis system is secure and trustworthy. Privacy-preserved data-leak detection can be achieved by leveraging special protocols and computation steps. It is another functionality of a detection system.
- We implement the web service to maintain the users and sensible content instead of data bases because of static implementation and rough data handling. Even the sensible data storage have to preserved from threatens in existing system. For that purpose we used to maintain the sensible data in cloud.

5. MODULES:

1. Content Outsourcing without DLD.
2. Build data leakage detection framework.
3. Content Outsourcing with DLD Checker.
4. Sensitive Data Detection

1. *Content Outsourcing without DLD:*

In this module user's register in mail server with their name, authorized job position and their authorized e-mail domain. And the users can transfer their file using without any restriction of sensible content checking.
There is no content checking and domain filtering on their transformed sensible data. Sensible content is outsourcing from one organization to another organization performed by user. The content can be of any file (text, document). Outsourcing will not reach DLD and directly reach its destination or organization. Here outsourcing mechanism of transferred data is offending over the protocol.

2. *Build data leakage detection framework:*

In this module mail server data owner generates a sensitive data and stored in the cloud and create the directory for lucene search framework and other data leakage detectors. Data owner's cloud contains much sensitive information about their authorized customer's details, information technology source, and database and server details. This sensitive information is maintained by Data Leak Detector. Using this DLD referenced directory perform data leak detection mechanism.

The DLD consist of lucene search engine framework, levenshtein distance algorithm and our own shuffled checking algorithm. The DLD directly configured with cloud and can refer every data transformation outsourcing from authorized user transformation.

3. *Content Outsourcing with DLD Checker:*

DLD is the one will check all the outsourcing content before it transmit to the other organization. All the outsourced contents are check with sensitive data. All the sensitive data are maintaining in index file. Using this index file DLD identify the sensitive data concurrently with domain filtering and threshold assigning based on their email domain. DLD will check every line of the sending data with the sensitive file. DLD will not allow any sensitive data will leak to any of the other organization.

In proxy mail server the every occurrence of transformed contents are filter by users email domain. All users' details are retrieved from the cloud using their email. Then threshold assigned for the users based on their authorized job position and the transferred content has been tested by lucene framework search engine, levenshtein distance checking and shuffling algorithm.

4. *Sensitive Data Detection:*

Once the DLD framework checks the outsourced content, if any data leak is identified means DLD will detect the sensitive data. Here DLD will check not only the sensitive data and also it will check some access condition. Every data owner maintain common access condition every file. For example, all the contents are encrypted before they outsourced. If DLD identified any sensitive information outsourcing means they will detect the sensible content in between of the file outsourcing.

For the purpose of false alert, we maintain threshold of every domain and users position. If the sensible content percentage of transferred file exceeds the

threshold percentage which trigger alert mail to Admin of the proxy mail server. Alert mail consists of entire details about the users even what are the sensible contents are pings from the transferred content by the DLD framework.

I. Software Requirements:

- Windows XP and Above
- JDK 1.7
- Tomcat 6.0

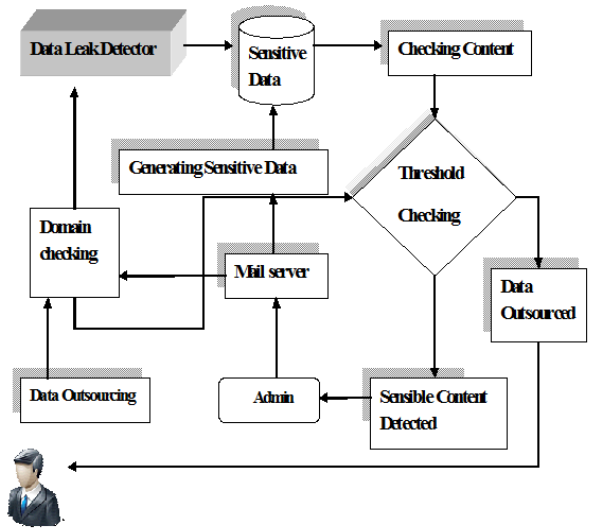
II. Hardware Requirements:

- Hard Disk : 40GB and Above
- RAM : 2GB and Above
- Processor : P IV and Above

III. Technologies Used:

- J2SE, J2EE (JSP, Servlets).
- Jersey Restful web service, Lucene framework, Json.
- JavaScript , HTML ,CSS

6. ARCHITECTURE DIAGRAM:



USER

7. CONCLUSION:

Hence we proposed and developed fast detection of data-leakage framework to avoid sensitive data exposure and also provide privacy-preserving to sensitive data. We presented a content inspection technique for detecting leaks of sensitive information in the content of files or network traffic. Our detection approach is based on aligning two sampled sequences for similarity comparison. Our experimental results suggest that our alignment method is useful for detecting multiple common data leak scenarios. The parallel versions of our prototype provide substantial speedup and indicate high scalability of our design. For future work, we plan to explore data-movement tracking approaches for data leak prevention on a host.

REFERENCES:

[1] X. Shu, J. Zhang, D. Yao, and W.-C. Feng, "Rapid and parallel content screening for detecting transformed data exposure," in Proc. 3rd Int. Workshop Secur. Privacy Big Data (BigSecurity), Apr./May 2015, pp. 191–196.

[2] X. Shu, J. Zhang, D. Yao, and W.-C. Feng, "Rapid screening of transformed data leaks with efficient algorithms and parallel computing," in Proc. 5th ACM Conf. Data Appl. Secur. Privacy (CODASPY), San Antonio, TX, USA, Mar. 2015, pp. 147–149.

[3] (Feb. 2015). Data Breach QuickView: 2014 Data Breach Trends. [Online]. Available: https://www.riskbasedsecurity.com/reports/2014-YEDataBreachQuickView.pdf, accessed Feb. 2015.

[4] Kaspersky Lab. (2014). Global Corporate IT Security Risks. [Online]. Available: http://media.kaspersky.com/en/business-security/Kaspersky_Global_IT_Security_Risks_Survey_report_Eng_final.pdf

[5] L. De Carli, R. Sommer, and S. Jha, "Beyond pattern matching: A concurrency model for stateful deep packet inspection," in Proc. ACM SIGSAC Conf. Comput. Commun. Secur., 2014, pp. 1378–1390.

[6] A. V. Aho and M. J. Corasick, "Efficient string matching: An aid to bibliographic search," Commun. ACM, vol. 18, no. 6, pp. 333–340, Jun. 1975.

[7] R. S. Boyer and J. S. Moore, "A fast string searching algorithm," Commun. ACM, vol. 20, no. 10, pp. 762–772, Oct. 1977.

[8] S. Kumar, B. Chandrasekaran, J. Turner, and G. Varghese, "Curing regular expressions matching algorithms from insomnia, amnesia, and acalculia," in Proc. 3rd ACM/IEEE Symp. Archit. Netw. Commun. Syst. (ANCS), 2007, pp. 155–164.

[9] H. A. Kholidy, F. Baiardi, and S. Hariri, "DDSGA: A data-driven semi-global alignment approach for detecting masquerade attacks," IEEE Trans. Dependable Secure Comput., vol. 12, no. 2, pp. 164–178, Mar./Apr. 2015.

[10] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," J. Molecular Biol., vol. 215, no. 3, pp. 403–410, Oct. 1990.