IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 4 Issue 2, February 2017
ISSN (Online) 2348 – 7968 | Impact Factor (2016) – 5.264
www.ijiset.com

Preserving Data from Leakage by Using Data Detection and Prevention Methods

Ms. A.Punitha,(Ph.D) [#1], Dr.V.Geetha,Ph.D[#2], Dr.A.Meiappane,Ph.D[#3]

[#1]Assistant Professor, [#3] Associate Professor
Department of Information Technology
Manakula Vinayagar Institute of Technology, Puducherry.

[#2]Associate Professor,
Department of Information Technology
Pondicherry Engineering College, Puducherry.

***Abstract*— Numbers of leaked sensitive data incidents are posse's serious threat to organizational security. Lack of proper encryption and communications on files by humans, leads to data loss. So organization needs tools 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, because of its unpredictable leak patterns. In this Paper work, we explore data *leak* method, by comparable sample algorithm & sampling oblivious alignment algorithm and discussed about the data prevention laws and technology, where we can keep the sensitive data locked. Hence these methods detect and prevent the organizational properties that are too sensitive from the unauthorized hands.**

***Index Terms*— sensitive data, detection, sampling alignment, Prevention.**

I. INTRODUCTION

Network security is protection of the access to files and directories in computer network against hacking, misuse and unauthorized changes to the system. An example of network security is an anti virus system. Network security starts with authenticating, commonly with a username and a password. Since this requires just one detail authenticating the user name i.e., the password this is sometimes termed one-factor authentication. With two-factor authentication, something the user is also used (e.g., a security token or 'dongle', an ATM card, or a mobile phone); and with three-factor authentication, something the user is also used (e.g., a fingerprint or retinal scan).

Once authenticated, a firewall enforces access policies such as what services are allowed to be accessed by the network users. Though effective to prevent unauthorized access, this component may fail to check potentially harmful content such as computer worms or Trojans being transmitted over the network. Anti-virus software or an intrusion prevention system (IPS) help detect and inhibit the action of such malware. An anomaly-based intrusion detection system may also monitor the network like wireshark traffic and may be logged for audit purposes and for later high-level analysis. Communication between two hosts using a network may be encrypted to maintain privacy.

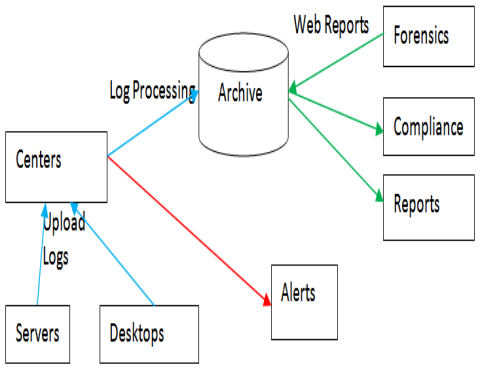

Fig: Data Leakage Representation

Reports show that the number of leaked sensitive data records has grown 10 times in t he last 4 years, and it reached a record high of 1.1 billion in 2014. A significant portion of the data leak incidents are due to human errors, for example, a lost or stolen laptop containing unencrypted sensitive files, or transmitting sensitive data without using end to- end encryption such as PGP. A recent Kaspersky Lab survey shows that accidental leak by staff is the leading cause for internal data leaks in corporate. The data-leak risks posed by accidents exceed the risks posed by vulnerable software.

In order to minimize the exposure of sensitive data and documents, an organization needs to prevent clear text sensitive data from appearing in the storage or communication. A screening tool can be deployed to scan computer file systems, server storage, and inspect outbound network traffic. Tool searches for the occurrences of plaintext sensitive data in the content of files or network traffic. It alerts users and administrators of the identified data exposure vulnerabilities. For example, an organization's mail server can inspect the content of outbound email

messages searching for sensitive data appearing in unencrypted messages.

II. RELATED WORK

The leak of sensitive data on computer systems poses a serious threat to organizational security. Organizations need 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. Existing automata-based string matching algorithms[4] are impractical for detecting transformed data leaks because of its formidable complexity when modelling the required regular expressions[3].

The importance of network security has grown tremendously and a collection of devices have been introduced, which can improve the security of a network. Network intrusion detection systems [5] are among the most widely deployed such system. Deterministic Finite Automata (DFA) are fast, therefore they are often desirable at high network link rates. DFA for the signatures [3], which are used in the current security devices, however require prohibitive amounts of memory, which limits their practical use.

III. OUR MODEL

Our proposal is to detect when the sensitive data has been leaked by agents, and if possible to identify the agent that leaked the data. Perturbation is a very useful technique where the data is modified and made "less sensitive" before being handed to agents. We introduce the sampling alignment algorithm where the data is sequenced and content being inspected to detecting leakage of a set of objects or records, this Algorithm achieves sampling oblivion by inferring the similarity between null regions (consecutive sampled-out elements) and counts that similarity in the overall comparison outcomes between the two sampled sequences. In this way that improves our chances of identifying a leaker. After the work of sensitive data leakage identified, the prevention technology is applied, is intended to stop unauthorized users from sending sensitive or critical information outside of the corporate network, this improves the data security of the organization.

A. Data Leakage Detection

The leak of sensitive data on computer systems poses a serious threat to organizational security due to human error. Organizations need to detect the leakage of data in effective manner. Here Data leak detection approach introduced by providing the pair of a comparable sam-pling algorithm and a sampling-oblivious alignment algorithm. The comparable sampling algorithm yields constant samples of a sequence wherever the sampling starts and ends. The sampling-oblivious alignment algorithm infers the similarity between the original unsampled sequences with sophisticated trackback techniques through dynamic programming.

(1) Comparable Sampling Algorithm

We present our comparable sampling algorithm; the selection decision of an item depends on how it compares with its surrounding items according to a selection function. As a result, the sampling algorithm is deterministic and subsequence-preserving.

Algorithm:

Input: an array S of items, a size $|w|$ for a sliding window w, a selection function $f(w, N)$ that selects N smallest items from a window w, i.e., $f = \min(w, N)$

Output: a sampled array T

initialize T as an empty array of size $|S|$
$w \leftarrow$ read$(S, |w|)$
let $w.head$ and $w.tail$ be indices in S corresponding to the higher-indexed end and lower-indexed end of w, respectively
collection mc ← min(w, N)
while w is within the boundary of S do
m p ← mc
move w toward high index by 1
mc ← min(w, N)
if mc = m p then
item en ← collectionDiff (mc, m p)
item eo ← collectionDiff(m p , mc)
if en < eo then
write value en to T at w.head's position
else
write value eo to T at w.t ail's position
end if
end if
end while

Sampling Algorithm Analysis: this sampling algorithm is deterministic and same time it satisfies the subsequence-preserving requirement. i.e., given a fixed selection function f : same inputs yield the same sampled string. However, deterministic sampling does not necessarily imply subsequence preserving. One can prove using a counterexample. Consider a sampling method that selects the first of every 10 items from a sequence, e.g., 1-st, 11-th, 21-st, It is deterministic and subsequent preserving one.

(2) Sampling-oblivious alignment

We observe that leaked data region is usually consecutive and it has more number of null regions in it. Thus, our algorithm achieves sampling oblivion by inferring the similarity between null regions and counts that similarity of sampled sequences. The inference is based on the comparison outcomes between items surrounding null regions and sizes of null regions. For example, given two sampled sequences a-b and A-B, if a == A and b == B, then the two values in the positions of the null regions are likely to match as well.

Example: Sampling-oblivious alignment

```
Original lists:
        562798385743254639782436 6
          56279839664325463 95
Sampled sequences need to be aligned
as:

--2---3-5---2---3-7-2-3--

--2---3-6---2---3–
```

Sampling oblivious alignment Algorithm Analysis: The complexity of our alignment algorithm produce lengths of compact and the alignment complexity for a single piece of sensitive data of size l is the same as that of a set of shorter pieces.

Table I

Evaluation on Detection Accuracy

Data Prevention	Security & Confident ability	Rules & Regulation	trustworthy & educated personnel	Well Defined	Maintenance of Technology	Third Party Proxies
Laws	✓	✓	✓			
Technology				✓	✓	✓

B. Data Leakage Prevention

Data leak prevention (DLP) is a set of information security tools which is used to stop users from sending sensitive or critical information outside network. Information loss prevention, which is done by insider threats and it, can be prevented adapting various privacy laws and Prevention Technologies.

(1) The laws emphases on Security and Confident ability, which define the security risks to the best of one's knowledge, appropriate security measures based on the risk estimation and comply or do not comply with regulations. Organization also forces its employees to comply with these regulations by using control mechanisms, surveillance, and monitoring such as intrusion prevention systems and intrusion detection system. Even though security Risk can be reducing by trustworthy and educated personnel are required to implement, Configure, manage, monitor, and operate the technical installations within the infrastructure of an organization. Furthermore, it is important to keep their competence, training, and awareness up to date.

(2) Data leak prevention (DLP) Technology, which used for identifying well-defined content like credit cards numbers, leads to identify other sensitive data, like intellectual formulas. This successful implementation requires significant maintenance of technology, third-party proxies, mail-filtering to stop movement of internetwork and web-based traffic.

However, even if all levels such as laws and technology, considered and implemented there is a security risk which is compromised.

Table II

Evaluation on Prevention Accuracy

Category	Preserving	Deterministic	Accuracy	Inference	Complexity
Data Detection	✓	✓	✓	✓	✓

IV. CONCLUSION:

We can provide security to our data during its distribution or transmission and even we can detect if that gets leaked. Thus, using this model security system is developed by using the sampling alignment algorithm. Data Prevention laws and technology is discussed. Finally, the obstacles in the data security are overcome in this paper by implementing the data leakage detection and following the Data prevention laws and technology.

REFERENCES

[1] Shu, Jing Zhang, Danfeng Yao, Senior Member and Wu-Chun Feng, "Fast Detection of Transformed Data Leaks" in IEEE Transactions on information forensics and security, vol. 11, no.3, March 2016.

[2] Barbara Hauer, "Data and Information Leakage Prevention within the Scope of Information Security", in IEEE Transactions on information forensics and security, Digital Object Identifier 10.1109/ACCESS.2015.2506185, December 7, 2015.

[3] 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.

[4] 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.

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

[6] 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

[7] 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.

[8] 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.

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

[10] 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