ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

Sensitive Data Leak Detection System using Fingerprint of Data

Aruna J.[1], Sabarinathan P.[2]

PG scholar, Dept. of CSE, Pavendar Bharathidasan College of Engineering and Technology, Trichy, Tamilnadu, India[1]

Assistant professor, Dept. of CSE, Pavendar Bharathidasan College of Engineering and Technology, Trichy, Tamilnadu, India[2]

Abstract—Data leak is the common issue of the computer system. To present a privacy preserving Data Leak Detection (DLD) method to solve this problem where a set of data is leaked. The advantage of this detection method is that it enables the data owner to safely delegate the detection process without fully exposing the sensitive data to a semi-honest provider. DLD service is offered to the user with strong privacy guarantees. The proposed system is used to generate the fingerprint and generate the token id for the sensitive data and then check it with the sensitive data. The token id is available in the database. To propose an improvement for this approach to offer a much faster processing time with accuracy. The core idea for this system is to remove the types of phrases from the fingerprinting process. This types of phrases are identified by looking at public documents of the organization that want to protect from the data leaks.

Keywords—sensitive data, data leak, token id, accuracy, privacy.

I. INTRODUCTION

Data leakage is the big challenge of an organizations. The confidential data leaks, it may be either accidental or intentional and it may cause huge data losses to the data owner.Network data-leak-detection is a method, it performs deep packet inspection (DPI) over a network channel. DPI is used to analyze the TCP/IP packets for inspecting the data, when the data found in network traffic then give alerts to the organization. If the detection system is outsourced then it may expose the sensitive data to the unauthorized user. To propose the fingerprint algorithm to solve this problem that enhances data privacy during the process. This approach is based on the one-way computation. It can support the data owner to safely delegate the detection operation without exposing the sensitive data. In this detection operation the data owner to prepare the fingerprint and then release the fingerprint, small amount of data to the DLD provider. Data owner does not want to directly expose the sensitive data to the provider. The DLD provider continuously monitor the network channel and check any data leaks are found over a channel. If any leaks are found immediately send all data leak reports to the data owner. Now the data owner can decide whether or not it is a data leak also identifying the guilt agents. During the monitoring process the DLD provider gain exact knowledge about the sensitive data. The security goal of this method is to detect the inadvertent data leaks caused by human mistakes. The privacy goal of the fuzzy fingerprint mechanism to prevent the DLD provider from gaining the exact value about the data during the operation. It means that the DLD provider given digests of the sensitive data to the owner then the content of the network traffic to be examined. The DLD provider should not learn the exact value of the sensitive data. The privacy policy model is used to protect the sensitive data from the privacy violations. The fuzzy fingerprint technique is used to hide the sensitive data in network traffic, it prevents the DLD provider as it learns the content of the sensitive data.

The sensitive data is accidentally leaked in the outbound traffic by a legitimate user. To focuses on detecting this type of accidental data leaks over supervised network channels. Inadvertent data leakmay be due to

human errors such as forgetting to use encryption, carelessly forwarding an internal email and attachments to outsiders, or due to application flaws. A supervised network channel could be an unencrypted channelor 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 (manin-the-middle) SSL sessions are established instead of normal SSL.

To provide a solution based on data retrieval to identify the phrases containing in sensitive data for fingerprinting. The main goal of this system is to identify the popularity of phrases before starting the fingerprinting. This solution can support to improve the accuracy of detection operation and reducing the false positives, even when the sensitive data has been transformed in a network traffic. False positives caused by human and common phrases. To quantifying the data leak capacity in outbound network traffic, it is used to calculate the false positive rate during the detection operation.

II. RELATED WORK

There are several advances in security applications, it gives more security to the sensitive data. The fuzzy fingerprint mechanism to identify the outsourced DLD server and provide a systematic solution to this problem. There existing system, shingle and Rabin fingerprint technique was used for identifying the data leaks in a collaborative setting. To propose the fuzzy fingerprint algorithm gives the privacy preserving data leak detection solution with convincing results. Most data leak detection products do not have the privacy preserving feature and this products are offered by the industries. The proposed system approach is different from the other approach and it can provides the data leak detection service. Using this method the data owner does not need to fully reveal the sensitive data to the DLD provider.

Bloom filter is used in the network security layers from network security to application security, it is a space-saving data structure for se membership test. The fuzzy Bloom filter invented to constructs a special Bloom filter, it sets the corresponding filter bits to1's. This method is a potential privacy preserving technique. The fuzzification process is used in fuzzy fingerprint technique, it is separated from the membership test, and it is flexible to test whether the fingerprint is sensitive with or without fuzzification. Privacy preserving keyword search or fuzzy keyword search provide string matching in semi honest environments. Anomaly detection can be used to detect data leaks in network traffic. It detects the new information in traffic, entropy analysis is used in this detection process. To present a signature based model to monitor the design can be outsourced also detect the data leaks. Both the anomaly detection and signature based detection approaches are different.

Tracing and enforcing are another approaches for data leak detection. It contains data flow and file-descriptor sharing enforcement. This approaches do not provide a remote service so this approaches are different from ours. The fuzzy fingerprint approach some other privacy preserving methods are invented for specific process, e.g., secure multi-party computation. SMC is a cryptographic mechanism it supports the string matching also complex functions. The advantage of the proposed system is its concision and efficiency.However, one issue of this approach is that true positive and false positives yield the same fingerprint value due to collision, which prevents the data owner from telling true positives apart from false positives. In addition, our fuzzy fingerprint approach is more flexible from the deployment perspective, as the data owner can adjust and fine-tune the privacy and accuracy in the detection without recomputing the

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)
Vol. 2, Special Issue 10, March 2016

fingerprints. In contrast, the precision is fixed in the naive shorter polynomial approach unless fingerprints are recomputed. An efficient technique to address this problem. The main idea is to relax the comparison criteria by strategically introducing matching instances on the DLD provider's side without increasing false alarms for the data owner. Specifically, *i)* the data owner perturbs the sensitive-data fingerprints before disclosing them to the DLD provider, and *ii)* the DLD provider detects leaking by a range-based comparison instead of the exact match. The range used in the comparison is pre-defined by the data owner and correlates to the perturbation procedure. The data owner to prepare the digests of sensitive data, release for the data owner to send the digests to the DLD provider, monitor and detect for the DLD provider to collect outgoing traffic of the organization, compute digests of traffic content, and identify potential leaks, report for the DLD provider to return data-leak alerts to the data owner where there may be false positives (i.e., false alarms). This is based on strategically computing data similarity, specifically the quantitative similarity between the sensitive information and the observed network traffic. High similarity indicates potential data leak. For data-leak detection, the ability to tolerate a certain degree of data transformation in traffic is important.

The proposed system, an information retrieval based solution is used to improve the performance of the cyclical hashing method for the sensitive data leak detection (IRILD). During the fingerprinting process to identify and remove the pubic phrases and common phrases from this process. Pubic phrases found in the public documents and common phrases identified by checking the number of results returned by website when querying the phrases since this types of phrases not contain sensitive data. Specifically IRILD achieved a much faster data leak detection process speed also achieved higher accuracy compared with cyclical hashing. The competitive benefits of developing a "one-stop-shop" silver bullet data leakage detection to provide the highest degree of protection by ensuring an optimal fit of specific data leakage detection technologies with the threat landscape. This landscape is characterized by the types of leakage channels. Data leakage detection is performed by using watermarking. It can be very useful in data leakage detection method, sometimes it can destroyed if the data recipient is malicious. The benefit of agent-based information leakage detection system is used to modify and add detection capabilities. Mobile agents to provide unique reporting capabilities that can be used to calculate the data leak detection accuracy. Data plays an important role in IT system that is when the data has to be sent to other user through the trusted agents and it is very challenging. Where the distributor provide the sensitive data to the trusted parties and the data is intentionally leaked to others. The distributor should identify or detect this leakage and its means that is who leaked it as well.

III. PROPOSED SYSTEM

Data leakage is the big challenge in front of the industries & different institutes. Though there are number of systems designed for the data security by using different encryption algorithms, there is a big issue of the integrity of the users of those systems. It is very hard for any system administrator to trace out the data leaker among the system users. It creates a lot many ethical issues in the working environment of the office.

The data leakage detection industry is very heterogeneous as it evolved out of ripe product lines of leading IT security vendors. A broad arsenal of enabling technologies such as firewalls, encryption, access control, identity management, machine learning content/context based detectors and others have already been

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)
Vol. 2, Special Issue 10, March 2016

incorporated to offer protection against various facets of the data leakage threat. The solution for this problem is to generate the fingerprint for sensitive data and generate the token id for the fingerprints.

A. Design Considerations:

- Generate fingerprint for each sensitive data.
- Generate the token id for sensitive data.
- Release the fingerprint and reveal the small amount of data to the provider.
- DLD provider monitor the network traffic.
- Detect the data leaks.
- Report all data leak alerts to the data owner, it enables to identify the guilt agents.
- Data owner decide whether or not it is a true leak.

B. Description of the Proposed Algorithm:

The main goal of the proposed algorithm is to discover the appearance of the sensitive data over a supervised network channel and prevents the DLD provider to learn the content of the data.

The major contributions of the proposed system is:

- To propose a solution to improve the performance of the traditional approach for information leak detection. The main idea is to identify non-sensitive phrases as well as common phrases, and remove them from the fingerprinting process of confidential documents.
- To evaluate the popularity of a long combined phrase to provide a technique to split the phrase into sub-phrases and finding out the popularity of the phrase based on its divided phrases.

C. *Limitations:*

To analyze the security and privacy guarantees provided by our data-leak detection system, as well as discuss the sources of possible false negatives data leak cases being overlooked and false positives legitimate traffic misclassified as data leak in the detection. To point out the limitations associated with the proposed network-based DLD approaches.

There are three limitations are used in this method:

Modified data leak, the shingle has the limited power to detect fully modified data leaks. The data is modified the data leak detection failure may occur. Advanced content comparison is needed to solve this issue. Dynamic sensitive data, it is used to protect the dynamically changing data. The digests continuously need to update. Raise question to the community for this problem. Selective fragments leak, false negative may occur using the subset of the sensitive data scheme (partial disclosure).

IV. SYSTEM DESIGN

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

A. *System Design:*

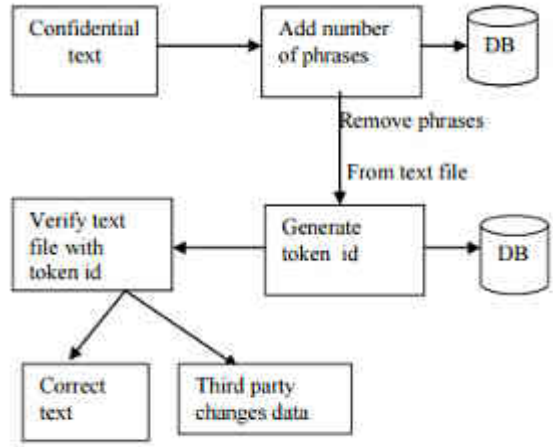

Figure 1: System design for Information leak Detection System

This method generates fingerprints for confidential documents in three steps. First we take input text file and then secondly we generate token ID for the specific text file by removing and third steps is to verify this text file If the result is true then text is correct and if the result is false then we conclude that someone tampered with this text.

V. SYSTEM IMPLEMENTATION

To avoid false positives involving phrases as shown in the example of Figure 1 and also reduce the un-necessary cost of generating and checking fingerprint of the phrases from database, to propose fingerprint based method that is able to identify phrases and eliminate them from the fingerprinting process. In this method, we evaluate the popularity of phrases by submitting them to in database that contains large number of phrases. Information leak detection system and also eliminate phrases that can be found in those public documents from the fingerprinting process because these phrases contain already known information.

To generates fingerprints for confidential documents in three steps:

- First we take input text file
- Secondly we generate token ID for the specific text file by removing and
- Third steps is to verify this text file as illustrated infig. 1.

While the fingerprint generation of information leak detection system is different from that of the popular approach, the information leak detection of these two approaches is still the same, i.e., fingerprints of confidential documents in the database are used to check against fingerprints of outgoing documents for information leak detection. This method is not used for encoded, encrypted, or compressed data. The DLD provider obtains digests from the data owner for each sensitive data.

To calculate the accuracy first test the detection rate and false positive rate and check where the sensitive data is leaked or not leaked in original form. Simple leaking scenarios is used to test the prototype without partial disclosure. There are three experiments are performed in this scenarios. First one is true leak, the entire set of sensitive data is leaked via FTP. Next one is no leak, the DLD server analyze the network traffic

and confirm no data leaks are found. Last one is no leak, after that no sensitive data should be confirmed by the data owner. The first experiments is designed to the detection operation and last two is designed to estimate the false positive rate. The results show that the accuracy of the sensitive data. Complex leak scenarios, the data owner may partially expose the sensitive data's fingerprint to the DLD server for detection operation. This scenario used to measuring the percentage of the exposed sensitive data in traffic.

VI. CONCLUSION

The proposed algorithm performance is better with high accuracy and low false positive rate. The fingerprint method is used to identify the data leakage. It can support the accurate detection with low false positives.This proposed method is used for information leak detection is to generate fingerprint of the confidential document and generate token id that is available in Database and then check it with confidential document. In this paper to propose an improvement for this approach to offer a much faster processing time with accuracy. The core idea of our solution is to eliminate types of phrases from the fingerprinting process. Types of phrases are identified by looking at available public documents of the organization that we want to protect from information leaks and different phrases are identified with the help Databases.

REFERENCES

[1] X. Shu and D. D. Yao, "Data leak detection as a service," in *Proceedings of the 8th International Conference on Security and Privacy in Communication Networks*, pp. 222–240, 2012.

[2] Risk Based Security, "Data breach quickview: An executives guide to 2013 data breach trends," February 2014, https://www.riskbasedsecurity. com/reports/2013-DataBreachQuickView.pdf, accessed October 2014.

[3] Ponemon Institute, "2013 cost of data breach study: Global analysis," May 2013, https://www4.symantec.com/mktginfo/whitepaper/053013 GL NA WP Ponemon-2013-Cost-of-a-Data-Breach-Report daiNA cta72382.pdf, accessed October 2014.

[4] Identity Finder, "Discover sensitive data prevent breaches DLP data loss prevention," http://www.identityfinder.com/, accessed October 2014.

[5] K. Borders and A. Prakash, "Quantifying information leaks in outbound web traffic," in *Proceedings of the 30th IEEE Symposium on Security and Privacy*, pp. 129–140, 2009.

[6] H. Yin, D. Song, M. Egele, C. Kruegel, and E. Kirda, "Panorama: capturing system-wide information flow for malware detection and analysis," in *Proceedings of the 14th ACM conference on Computer and Communications Security*, 2007, pp. 116–127, 2007.

[7] K. Borders, E. V. Weele, B. Lau, and A. Prakash, "Protecting confidential data on personal computers with storage capsules," in *Proceedings of the 18th USENIX Security Symposium*, pp. 367–382, 2009.

[8] A. Nadkarni and W. Enck, "Preventing accidental data disclosure in modern operating systems," in *Proceedings of the 20th ACM conference on Computer and Communications Security*, 2013, pp. 1029–1042, 2013.

[9] A. Kapravelos, Y. Shoshitaishvili, M. Cova, C. Kruegel, and G. Vigna, "Revolver: An automated approach to the detection of evasive web-based malware," in *Proceedings of the 22nd USENIX Security Symposium*, 2013.

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

[10] X. Jiang, X. Wang, and D. Xu, "Stealthy malware detection and monitoring through VMM-based "out-of-the-box" semantic view reconstruction," *ACM Transactions on Information and System Security*, vol. 13, no. 2, p. 12, 2010.