ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 10, March 2016

HOST ASSISTED MECHANISM FOR DATA LEAK DETECTION

*Sumi.M, **Mrs.Bonshia Binu.M.S
Ponjesly college of Engineering
msumi022@gmail.com,binubonshia1988@yahoo.com

ABSTRACT-Number of data leaks in the organization, research institutions and security firms have grown rapidly in recent years. The data leakage occurs if there is no proper protection. The common approach is to monitor the data that are stored in the organization local network. The existing method considers all the data as sensitive and perform detection operation for all those data. However, this makes the detection process difficult and detection time to increase. In the proposed system, host assisted mechanism is used which takes only highly differentiated data as sensitive. Fuzzy fingerprint technique is designed and implemented that enhances data privacy during data-leak detection operations. The fingerprints generated are given to the DLD provider to perform operation. The DLD provider will perform operation for the sensitive data fingerprints and find out the leakages quickly. The evaluation results show that this method can provide accurate detection.

Keywords- Host assisted, data-leak, fingerprint, sensitive data, network traffic.

1. INTRODUCTION

The number of leaked sensitive data records has increased dramatically 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. Deep Packet Inspection is a technique to analyze the payloads of IP/TCP packets for inspecting application layer data, e.g., HTTP header/content. 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 existing network intrusion detection systems (NIDS).

Straightforward 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 memory). The existing method makes the data owner to consider all the data as sensitive and generate fingerprints for those data. Then the provider with the help of those fingerprints find whether there will be any chance of leakages in those data. This method takes more time for detection.

In the proposed system, host assisted mechanism is used which generates an automatic table containing the sensitive data. Automatic table is generated using statistical approach. Highly differentiated values are considered as sensitive and fingerprints are generated for them. Repeated values are ignored in this method. A data-leak detection solution is used which can be outsourced and be deployed in a semi-honest detection environment. Fuzzy fingerprint technique is implemented that enhances data privacy during data-leak detection operations. This approach is based on a fast and practical one-way computation on the sensitive data (SSN records, classified documents, sensitive emails, etc.). Using this detection method, the DLD provider, who is modeled as an honest-but-curious (aka semi- honest) adversary, can only gain limited knowledge about the sensitive data from either the released digests, or the content being inspected. Using this techniques, an Internet service provider (ISP) can perform detection on its customers' traffic securely and provide data leak detection as an add-on service for its customers.

2. EXISTING SYSTEM

The data-leak detection will perform detection operation for each and every data. Each and every data are considered as sensitive and fingerprints are generated for them. The existing method is semi automatic. It includes repeated values also and so repeated fingerprints are involved. Thus detection operation is done repeatedly for the same data. This makes the detection time to increase.

3. PROPOSED SYSTEM

In the proposed system, host assisted mechanism is used which checks the frequency of occurrence of data. Highly differentiated values are considered as sensitive and fingerprints are generated for them. Repeated values are ignored in this method. An automatic table containing the sensitive data is generated using statistical approach. The DLD provider computes fingerprints from network traffic and identifies potential leaks in them. To prevent the DLD provider from gathering exact knowledge about the sensitive data, the collection of potential leaks is composed of real leaks and noises. It is the data owner, who post-processes the potential leaks sent back by the DLD provider and determines whether there is any real data leak.

4. SYSTEM OVERVIEW

The privacy-preserving data-leak detection problem with a threat model, a security goal and a privacy goal is abstracted. First we describe the two most important players in the abstract model: the organization (i.e., data owner) and the data-leak detection (DLD) provider.

•**Organization** owns the sensitive data and authorizes the DLD provider to inspect the network traffic from the organizational networks for anomalies, namely inadvertent data leak. However, the organization does not want to directly reveal the sensitive data to the provider.

•**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.

Fig.1. System Overview

5. MODULE DESCRIPTION

Preprocess

The data owner stores their sensitive data in their network. They need their data to be in a protected way. They can't able to check the data frequently. Data leaks in organization, research institutions and security firms have grown rapidly in recent years. So the data owner delegate the detection operation semi-honest provider without revealing the sensitive data.

Host assisted mechanism

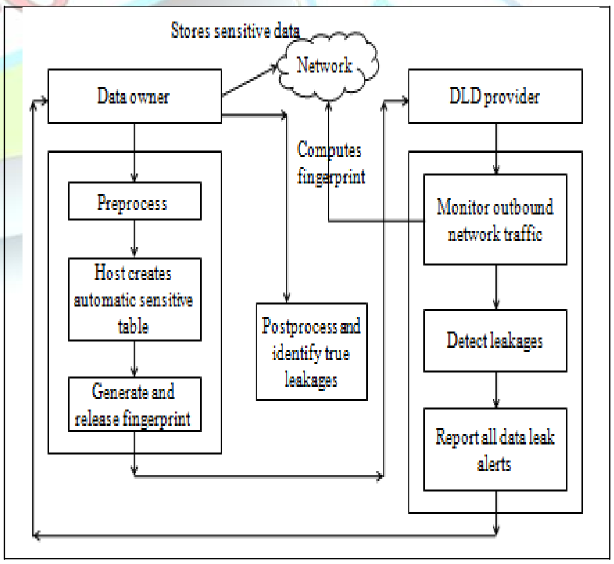
Host assisted mechanism is used after the data have been stored in the network. This checks the frequency of occurrence of data. Highly differentiated values are considered as sensitive and fingerprints are generated for them. Repeated values are ignored in this method. An automatic table containing the sensitive data is generated using statistical approach. Only the important data are considered here so the detection will be very effective through this mechanism.

Generate and release fingerprint

The data owner uses a sliding window and Rabin fingerprint algorithm to generate short and hard-to-reverse digests through the fast polynomial modulus operation. The sliding window generates small fragments of the processed data, which preserves the local features of the data and provides the noise tolerance property. Rabin fingerprints are computed as polynomial modulus operations, and can be implemented with fast XOR, shift, and table look- up operations. The Rabin fingerprint algorithm has a unique min-wise independence property, which supports fast random fingerprints selection for partial fingerprints disclosure.

A sliding window is used to generate q-grams on an input binary string first. The fingerprints of q-grams are then computed. A shingle is a fixed-size sequence of contiguous bytes. Local feature preservation is accomplished through the use of shingles. Therefore, this approach can tolerate sensitive data modification to some extent, e.g., inserted tags, small amount of character substitution, and lightly reformatted data. The use of shingles alone does not satisfy the one- wayness requirement. Rabin fingerprint is utilized to satisfy such requirement after shingling.

Monitor and detect leakages


Stores sensitive data
Network
Data owner
DLD provider
Computes fingerprint
Preprocess
Host creates automatic sensitive table
Generate and release fingerprint
Postprocess and identify true leakages
Monitor outbound network traffic
Detect leakages
Report all data leak alerts

The DLD provider obtains digests of sensitive data from the data owner. The DLD provider computes fingerprints from network traffic and identifies potential leaks in them. To prevent the

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 10, March 2016

DLD provider from gathering exact knowledge about the sensitive data, the collection of potential leaks is composed of real leaks and noises. If there is a data leak, then there will be a match between two fingerprints from sensitive data and network traffic. Then the DLD provider will send an alert to the data owner.

Report data leak alerts

It is the data owner, who post-processes the potential leaks sent back by the DLD provider and determines whether there is any real data leak. The data owner can tell whether a piece of sensitive data in the network traffic is a leak by using legitimate data transfer policies.

6. TECHNIQUE

Host assisted mechanism

Host assisted mechanism is used which takes only highly differentiated data as sensitive. Then fingerprints are generated for them. The DLD provider will perform operation for the sensitive data fingerprints and find out the leakages quickly

```
begin
          enter the data
          covert data into ASCII code
          view ASCII code
          find mean and standard deviation
          consider highly differentiated values as
          sensitive data
end
```

7. EVALUATION

The security and privacy guarantees provided by this data-leak detection system are analyzed. The limitations associated with the proposed network-based DLD approaches are pointed out.

A polynomial-time adversary has no greater than $2^{P_d - P_f}/n$ probability of correctly inferring a sensitive shingle, where P_d is the length of a fingerprint in bits, P_f is the fuzzy length, and $n \in [2^{P_d - P_f}, 2^{P_f}]$ 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 and appear in the traffic.

There is no match between sensitive and traffic fingerprints. The adversarial DLD provider needs to brute force reverse the Rabin fingerprinting computation to obtain the sensitive shingle. There are two difficulties in reversing a fingerprint: i) Rabin fingerprint is a one-way hash. ii) Multiple shingles can map to the same fingerprint. It requires to searching the complete set of possible shingles for a fingerprint and to identify the sensitive one from the set. This brute-force attack is difficult for a polynomial-time adversary, thus the success probability is not included.

7.2. Runtime comparison

Proposed system uses fingerprint filter implementation which is based on the Bloom filter library in Python (Pybloom). The runtime of Bloom filter provided by standard Pybloom (with dynamically selected hash function from MD5, SHA-1, SHA-256, SHA-384 and SHA-512) and that of fingerprint filter with Rabin fingerprint is compared. It shows that fingerprint filters run faster than Bloom filters, which is expected as Rabin fingerprint is easier to compute than MD5/SHA. The gap is not significant due to the fact that Python uses a virtualization architecture.

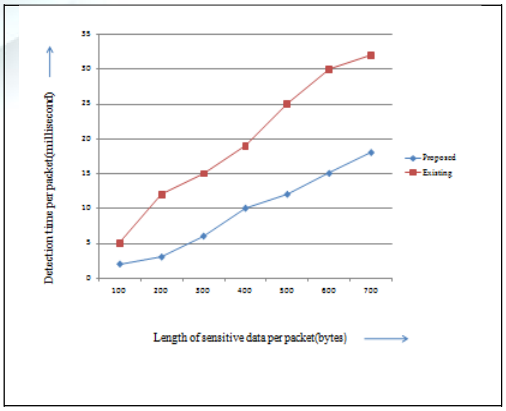


Fig.2.Detection time

8. RELATED WORK

In this paper, the privacy needs in an outsourced data-leak detection service and provide a systematic solution to enable privacy-preserving DLD services are identified. Shingle with Rabin fingerprint was used previously for identifying similar spam messages in a collaborative setting, as well as collaborative worm containment, virus scan, and fragment detection.

Most data-leak detection products offered by the industry, e.g., Symantec DLP, Global Velocity identity Finder, do not have the privacy-preserving feature and cannot be outsourced. GoCloudDLP is a little different, which allows its customers to outsource the detection to a fully honest DLD provider. Fuzzy fingerprint method differs from these solutions and enables its adopter to provide data-leak detection as a service. The customer or data owner does not need to fully trust the DLD provider using this approach.

Besides fuzzy fingerprint solution for data - leak detection, there are other privacy-preserving techniques invented for specific processes, e.g., DNA matching, or for general purpose use, e.g., secure multi-party computation (SMC). Similar to string matching methods discussed above, uses anonymous automata to perform comparison. SMC is a cryptographic mechanism, which supports a wide range of

fundamental arithmetic, set, and string operations as well as complex functions such as knapsack computation, automated trouble-shooting , network event statistics, private information retrieval genomic computation, private database query, private join operations, and distributed data mining. The provable privacy guarantees offered by SMC comes at a cost in terms of computational complexity and realization difficulty. The advantage of fuzzy fingerprint approach is its concision and efficiency.

9. CONCLUSION

This host assisted mechanism will help to find out the leakages quickly. Existing detection system takes all the data to conduct the detection operation. But in the proposed system the data owner the sensitive data is kept to a minimum level. It helps to delegate the detection operation to DLD provider without revealing sensitive data. The DLD provider can't able to get the shingle from the fingerprint as it is a one-way function. The privacy is achieved and detection operation is done efficiently. Detection time will be reduced through this mechanism.

REFERENCES

[1] Ananthi S., Sadish Sendil M. and Karthik S. (2011), 'Privacy preserving keyword search over encrypted cloud data', in Advances in Computing and Communications (Communications in Computer and Information Science), vol. 190. Berlin, Germany: Springer-Verlag , pp. 480–487.

[2] Elisa Costante, Sokratis Vavilis, Sandro Etalle, Jerry den Hartog, Milan Petkovic, and Nicola Zannone (2013), 'Database Anomalous Activities: Detection and Quantification', in SECRYPT, page 603- 608.

[3] Jason Croft and Matthew Caesar (2011), 'Towards practical avoidance of information leakage in enterprise networks', in Proc. 6th USENIX Conf. Hot Topics Secur. (HotSec), p. 7.

[4] Kangjie Lu, Zhichun Li, Vasileios Kemerlis, Zhenyu Wu, Long Lu, Cong Zheng, Zhiyun Qian, Wenke Lee, Guofei Jiang (2015), 'Checking More and Alerting Less: Detecting Privacy Leakages via Enhanced Data-flow Analysis and Peer Voting ', in Proceedings of the Network & Distributed System Security Symposium (NDSS), San Diego, CA.

[5] Kui Xu, Yao D., Ma Q., and Crowell A. (2011), 'Detecting infection onset with behavior-based policies', in Proc. 5th Int. Conf. Netw. Syst. Secur., pp. 57–64.

[6] Latha K., Nivedha P., Menagagandhi G., Ramya T. (2014), 'Collaborative Anti Spam Technique to Detect Spam Mails in E-Mails', in International Journal of Scientific and Research Publications, Volume 4, Issue 3,ISSN 2250-3153.

[7] Rashmi Bhat K. and Nikita Mahajan V. (2014), 'A Technique for Avoiding Data Leakage and Misuse',in IJARCSMS,Volume 2, Issue 2.

[8] Saranya S. and Srinivasan T.R. (2015), 'Secure Data Detection for Confidential Data Exposure', in IJIREST, Vol. III, Issue 04, Pp 47 – 54.

[9] Sneha Sahare, Mamta Joshi, Manish Gehlot (2012), 'Data Security in Local Networks Using Distributed Firewalls', in International Journal on Computer Science and Engineering (IJCSE), Vol. 4 No. 09.

[10] Xiaokui Shu and Elisa Bertino (2015), 'Privacy-preserving detection of sensitive data exposure', in IEEE trans.Information foren and Secur.,Vol.10, no.5.