ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

Privacy Preserving Sensitive Data Leakage Detection and Prevention

Aruna J, Sabarinathan P

PG Scholar, Dept. of CSE, Pavendar Bharathidasan College of Engineering and Technology, Trichy, India

Assistant Professor, Dept. of CSE, Pavendar Bharathidasan College of Engineering and Technology, Trichy, India

ABSTRACT: Data leakage is the common problem in the government or any other organizations and it has been growing rapidly, it is caused by human errors. The existing system to detect this type of data leaks and to provide alerts to that organization. The detection operation is secrecy in an existing system but it is difficult to satisfy in real time. During the detection process the detection server may be compromised. In proposed system to present a fuzzy fingerprint technique to solve this kinds of issues. The advantage of this technique is toimprove the data protection against the unauthorized data transmission, to provide network security to the sensitive data also identifying the guilt agents. The evaluation results indicate high accuracy, accurate detection with very low false alarms.

KEYWORDS- network security, sensitive data, data leak, privacy

I. INTRODUCTION

A network is a telecommunication network, it gives permission to the computer for data transmission. Data leakage is big challenge in an organization, there are several algorithms are designed for data security. 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 fuzzy 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 main goal of the fuzzy fingerprint technique is to hide the sensitive data in network traffic, it prevents the DLD provider as it learns the content of the sensitive data.

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

DOI: 10.15680/IJIRCCE.2016. 0401130

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. Thefuzzification 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.

III. PROPOSED ALGORITHM

A. *Design Considerations:*

- Generate fingerprint for each 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 proposed algorithm contains three main steps.

Step 1: Shingles and Fingerprints:

The DLD provider obtains digests from the data owner for each sensitive data. The data owner to generate the one-way computation digests using the shingle and Rabin fingerprint. A shingle is fixed-size sequence of the bytes. For example, 2-gram shingle set of string abcde consists of four elements (abbc cd de). The use of shingles stand alone does not satisfy the one-way computation requirements. After the shingling Rabin fingerprint is used to satisfy the one-way function requirements.

Step 2: Selection Criteria:

The fuzzy fingerprint is matched in network traffic, the DLD provider to detect the data leak and alerts are triggered then reports all data leaks to the data owner. The fuzzy fingerprint is not matched the DLD provider adversarial needs to reverse the Rabin fingerprinting computation for obtain the shingle.To quantify the alert rate in the network traffic for sensitive data.

The expected alert rate(R) is presented in eq. (1)

$$R = \frac{\alpha p_s K \tau}{n} = \frac{\alpha p_s \tau}{2^{p_f - p_d}} \qquad \text{eq. (1)}$$

Where $\tau$is the total number of fuzzified sensitive fingerprints,$n$is the expected traffic fingerprints set size, p_f is fingerprint length,$p_d$is fuzzy length, $p_s \in (0,1]$partial disclosure rate, and α the expected rate.

Step 3: Limitations:

There three limitations are used in this method. 1. 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. 2. 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. 3. Selective fragments leak, false negative may occur using the subset of the sensitive data scheme (partial disclosure).

IV. PSEUDO CODE

Step 1: Preprocess:
The data owner prepare the fingerprint for the sensitive data and chooses four parameters the length of a shingle (q), irreducible polynomial (p(x)), and fuzzy length is (p_d) and bitmask (M). The data owner computes the set (s) for all fingerprints of the sensitive data. The fingerprint is transforms into fuzzy fingerprint f^{*}. The data owner generates a random p_f-bit binary string for each fingerprint $\dot{f}$, mask out the bits not randomized by $\dot{f}'$ = (NOT M) AND $\dot{f}$ and fuzzify f with $f^{*}= f$ XOR $\dot{f}'$. The eq. (1) gives the result of this process.

$$f^{*}= ((\text{NOT } M) \text{ AND } \dot{f}) \text{ XOR } f \qquad \text{eq. (2)}$$

Step 2: Release:
This release process is performed by the data owner. The DLD provider collect the fuzzy fingerprint set S * from the data owner and the provider use this fuzzy set for detection operation.
Step 3: Monitor:
The DLD provider continuously monitor the outbound network traffic (T) from the data owner's organization. Each packet the payloads and Tis collected and send to the next process.
Step 4: Detection:
The DLD provider computes the set of Rabin fingerprint first. Then check where each fingerprint is also in S* using the fuzzy equivalence test eq. (2)

$$E(f', f^{*}) = \text{NOT } (M \text{ AND } (f' \text{ XOR } f^{*})) \qquad \text{eq. (3)}$$

Where f' is the true value that is original data leak.
The DLD provider aggregate the results from this step and raise alerts.
Step 5: Report:
If the data leaks alerts are triggered then the provider send reports to the data owner. The data owner test whether it is in set (s) and identify the true leak.

V. SIMULATION RESULTS

To implement the fuzzy fingerprint framework to detect the data leak and provide network security to the sensitive data. In all experiments to ensure the shingles are in 8-byte and the fingerprints is 32-bit. The sensitivity of the packet $S_{packet}$is used by the provider for detection operation. To evaluate the accuracy using the simple and complex leaking scenarios. 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. Fig. 1 shows the average sensitivity of each data packet.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. Fig. 2 shows the number of detected sensitive packets. The fingerprint filter method is based on the Bloom filter. This results contain both inbound and outbound network traffic of sensitive data.

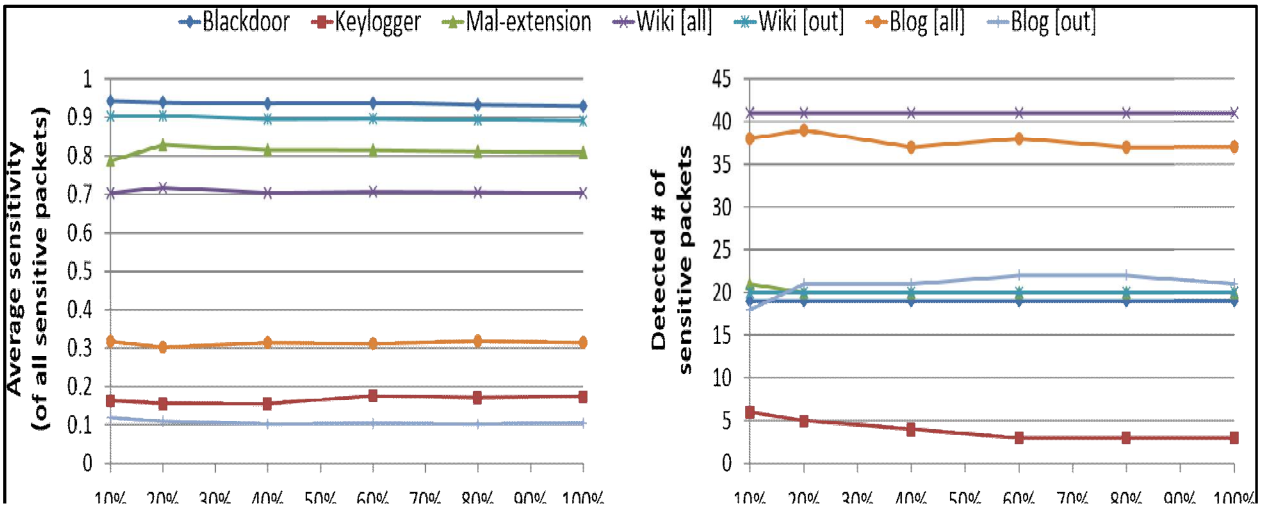

Fig. 1. The averaged sensitivity

Fig. 2. The number of detected sensitive packets

To compare the runtime of fingerprint filter with Rabin fingerprint and Bloom filter. To test their performance with hash functions. Fig. 3 shows the fingerprint filter is faster than Bloom filter. In Bloom filter the number of hash functions used it does not impact their runtime.

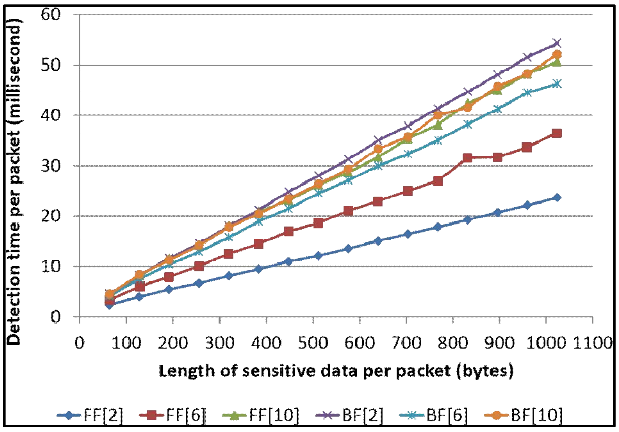

Fig. 3. Detection time

The fuzzy set size is corresponds to the K value based on the privacy goal. The fuzzy set size is based on the fuzzy length. The set of fuzzy size is small the set of fuzzy length is small. Fig. 4 shows the observed sizes of the fuzzy sets for fuzzy length. This experiment result is used to determine the fuzzy length of the dataset by the data owner.

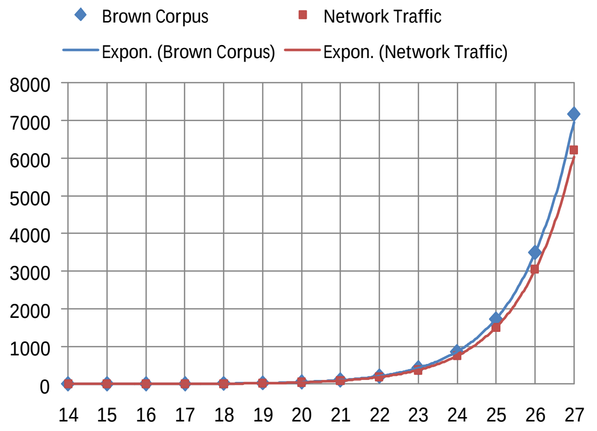

Fig. 4. The observed and expected sizes of fuzzy set for fuzzy length

VI. Conclusion and Future Work

The simulation resultsindicates the proposed algorithm performance is better with high accuracy and low false positive rate. The fuzzy fingerprint method is used to identify the data leakage. It can support the accurate detection with low false positives. For future work to focus on designing a host-assisted mechanism for complete data-leak detection and test their performance for large-scale organizations. To generate the token id that it is available in the database to check it with the sensitive data.

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. 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*, pp. 1029–1042, 2013.
3. B. Wang, S. Yu, W. Lou, and Y. T. Hou, “Privacy-preserving multikeyword fuzzy search over encrypted data in the cloud,” in *Proceedings of the 33th IEEE Conference on Computer Communications*, 2014, pp. 2112–2120.
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. 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*, 2009, pp. 367–382, 2009.

BIOGRAPHY

ARUNA J is a PG scholar Department of CSE in PavendarBarathidasan College of Engineering and Technology, Trichirappalli, Tamilnadu, India.

SABARINATHAN P is an Assiatant professor Department of CSE in PavendarBarathidasan College of Engineering and Technology, Trichirappalli, Tamilnadu, India.

DOI: 10.15680/IJIRCCE.2016. 0401130