ISSN 2350-1022

International Journal of Recent Research in Mathematics Computer Science and Information Technology
Vol. 2, Issue 2, pp: (121-126), Month: October 2015 – March 2016, Available at: www.paperpublications.org

Privacy Preserving Data Leak Detection for Sensitive Data

[1]Sumi.M, [2]Mrs.Bonshia Binu.M.S

[1]PG Student, CSE department, Ponjesly college of Engineering
[2]Assistant Professor, CSE department, Ponjesly college of Engineering

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 require the plaintext sensitive data. However, this requirement is undesirable, as it may threaten the confidentiality of the sensitive information. A privacy preserving data-leak detection solution is proposed which can be outsourced and be deployed in a semi-honest detection environment. Fuzzy fingerprint technique is designed and implemented that enhances data privacy during data-leak detection operations. 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. The evaluation results show that this method can provide accurate detection.**

Keywords: **Privacy preserving, 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 (DPI) is a technique to analyze analyze 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). In addition, the data owner may need to outsource the data-leak detection to providers, but may be unwilling to reveal the plaintext sensitive data to them. Therefore, one needs new data-leak detection solutions that allow the providers to scan content for leaks without learning the sensitive information.

In this paper, a data-leak detection solution is proposed 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. Our approach is based on a fast and practical one-way computation on the sensitive data (SSN records, classified documents, sensitive emails, etc.). It enables the data owner to securely delegate the content-inspection task to DLD providers without exposing the sensitive data. Using our 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 our 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.

ISSN 2350-1022

International Journal of Recent Research in Mathematics Computer Science and Information Technology
Vol. 2, Issue 2, pp: (121-126), Month: October 2015 – March 2016, Available at: www.paperpublications.org

2. EXISTING SYSTEM

The data-leak detection which used before 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). In addition, the data owner may need to outsource the data-leak detection to providers, but may be unwilling to reveal the plaintext sensitive data to them.

3. PROPOSED SYSTEM

A data-leak detection solution is proposed which can be outsourced and be deployed in a semi-honest detection environment. Fuzzy fingerprint technique is designed and implemented that enhances data privacy during data-leak detection operations. The data owner computes a special set of digests or fingerprints from the sensitive data and then discloses only a small amount of them to the DLD provider. 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 our 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.

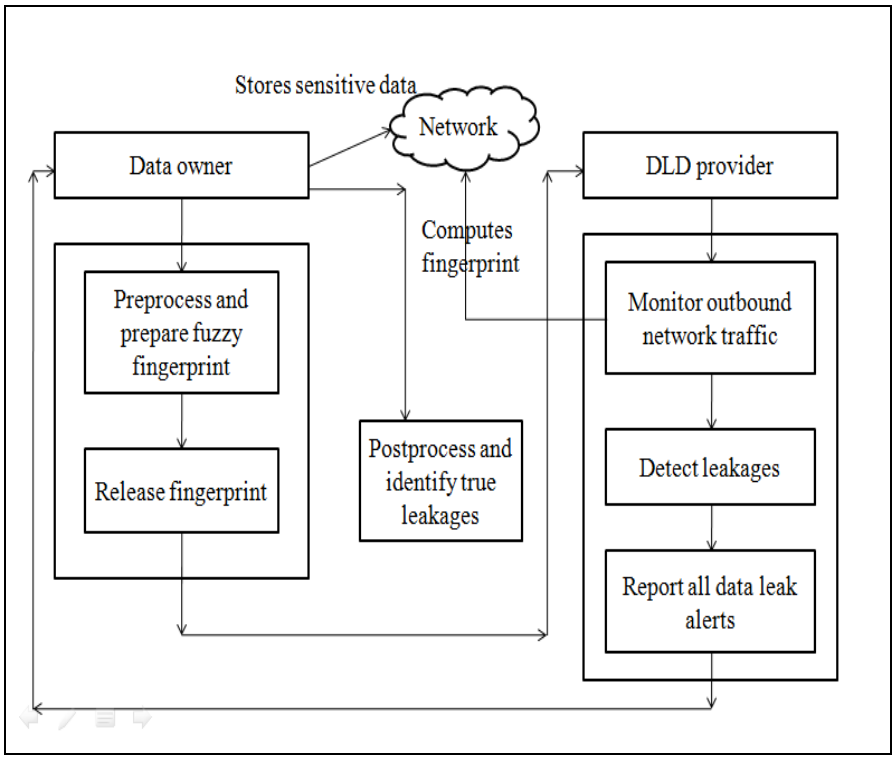

Fig.1: System Overview

ISSN 2350-1022

International Journal of Recent Research in Mathematics Computer Science and Information Technology
Vol. 2, Issue 2, pp: (121-126), Month: October 2015 – March 2016, Available at: www.paperpublications.org

5. MODULE DESCRIPTION

5.1. 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.

5.2. 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, our 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.

5.3. Monitor and detect leakages:

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 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.

5.4. 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. ALGORITHM

6.1. Rabin fingerprint algorithm:

This algorithm helps in generating a fuzzy fingerprint. It is used only after shingle is generated. generate digests of sensitive data through a one-way function, and then hide the sensitive values among other non-sensitive values via fuzzification. Using the min-wise independent property of Rabin fingerprint, the data owner can quickly disclose partial fuzzy fingerprints to the DLD provider. The purpose of partial disclosure is two-fold: i) to increase the scalability of the comparison in the DETECT operation, and ii) to reduce the exposure of data to the DLD provider for privacy.

```
begin
      int prime = 101;
   for each text
       for (int i = 0; i < text.Length; i++)
              char c = text[i];
              hash1=c*(int)(Math.Pow(prime, text.Length -
1 - i));
               fingerprint=hash1;
        end for;
   end for;
end;
```

ISSN 2350-1022

International Journal of Recent Research in Mathematics Computer Science and Information Technology
Vol. 2, Issue 2, pp: (121-126), Month: October 2015 – March 2016, Available at: www.paperpublications.org

7. EVALUATION

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

7.1. Privacy Analysis:

The privacy goal is to prevent the DLD provider from inferring the exact knowledge of all sensitive data, both the outsourced sensitive data and the matched digests in network traffic.

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.

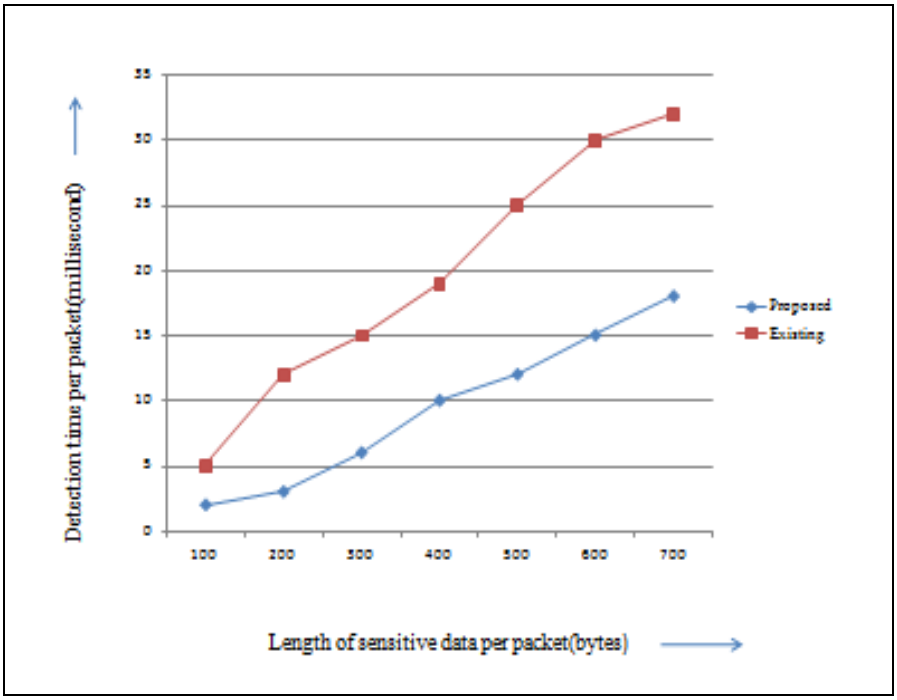

Fig.2: Detection time

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.

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 our 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 privacy preserving data-leak detection model make use of fuzzy fingerprint to find out the data leakages. Existing detection system requires the plaintext sensitive data to conduct the detection operation. But in the proposed system the data owner need not to give their sensitive data for finding data leakages. Instead fuzzy fingerprint is given. Using this model, the exposure of 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. For future work, a host-based mechanism is planned to be used in a large-scale organization.

REFERENCES

[1] K. Borders and A. Prakash, "Quantifying information leaks in outbound web traffic," in Proc. 30th IEEE Symp. Secur. Privacy, May 2009, pp. 129–140.

[2] G. Karjoth and M. Schunter, "A privacy policy model for enterprises,"in Proc. 15th IEEE Comput. Secur. Found. Workshop , Jun. 2002, pp. 271–281.

[3] Y. Jang, S. P. Chung, B. D. Payne, and W. Lee, "Gyrus: A framework for user-intent monitoring of text-based networked applications," in Proc.23rd USENIX Secur. Symp., 2014, pp. 79–93.

[4] K. Li, Z. Zhong, and L. Ramaswamy, "Privacy-aware collaborative spam filtering," IEEE Trans. Parallel Distrib. Syst., vol. 20, no. 5, pp. 725–739,May 2009.

[5] M. Cai, K. Hwang, Y.-K. Kwok, S.Song, and Y. Chen, "Collaborative Internet worm containment,"IEEE Security Privacy, vol. 3, no. 3, pp. 25–33, May 2005.

[6] J. Li, Q. Wang, C. Wang, N. Cao, K. Ren, and W. Lou, "Fuzzy keyword search over encrypted data in cloud computing," in Proc. 29th IEEE Conf. Comput. Commun., Mar. 2010, pp. 1–5.

[7] S. Ananthi, M. Sadish Sendil, and S. Karthik, "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, 2011, pp. 480–487.

[8] M. Burkhart, M. Strasser, D.Many, and X. Dimitropoulos, "SEPIA:Privacy-preserving aggregation of multi-domain network events and statistics," in Proc. 19th USENIX Conf. Secur. Symp., 2010, p. 15.

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

[10] X. Shu and D. Yao, "Data leak detection as a service," in Proc. 8th Int. Conf. Secur. Privacy Commun. Netw., 2012, pp. 222–240.