Secrecy Preserving Discovery of Subtle Statistic Contents

A.Francis Thivya
PG Scholar
Department of Computer Science & Engineering
Christian College of Engineering and Technology
Oddanchatram, Tamilnadu-624619, India

P.Tharcis
Assistant Professor
Department of Computer Science & Engineering
Christian College of Engineering and Technology
Oddanchatram, Tamilnadu-624619, India

Abstract

A Data Distributor or Software Company has given sensitive data to a set of confidential data owners. Sometimes data are leaked and found in the unauthorized place. Data leakage happens every day when the sensitive information is transferred to the third party agents. To manage and protect the confidential information, data leak detection is an important part. The exposure of sensitive information in storage and transmission places a serious threat to organizational and individual security. Data leak detection aims at scanning content of exposed sensitive information. Because of the large capacity and data volume, such a cryptographic algorithm needs to be scalable for a timely detection. In this project, the system proposes host based data leak detection (DLD). Data owner calculates a peculiar set of digests or fingerprints from the sensitive information, and then reveals only a small amount of digest data to the DLD provider. The system implements, and evaluates a new privacy-preserving data-leak detection system that changes the data owner to safely deploy locally, or to allocate the traffic-inspection task to DLD providers without revealing the sensitive data. It works well especially when consecutive data blocks are leaked. This host-based DLD technique may improve the accuracy, privacy, concision and efficiency of fuzzy fingerprint data leak detection.

Keywords: Data Leak Detection Provider, Fuzzy Fingerprint, Host-Based Data Leak Detection, Semi-Honest Adversary

I. INTRODUCTION

Information Security has always had an important role as technology has advanced; it has become one of the hottest topics of the last few decades. Information Forensic and Security is the investigation and analysis technique to gather and preserve information from a particular computer device. Information security is the set of business processes that protect information assets. It doesn't concentrate on how the information is formatted or produced.

Information security is very important for maintaining the availability, confidentiality and integrity of the information technology system and business data. Most of the large enterprises employ a dedicated security group to implement and manage the organizations information security program. Detecting and preventing data leak involves a set of solutions including data confinement [6], stealthy malware detection, policy enforcements and data leak detection. Typical approaches to preventing data leaks are under two categories [10] -- i) host-based solution and ii) network-based solution.

Network-based data leak detection focus on analyzing unscripted outbound network traffic through i) Deep packet inspection and ii) Information theoretic analysis. In deep packet inspection, inspecting every packet for the occurrence of sensitive data defined in database [1].

Network based data leak detection accompaniments host-based approach. Host based data leak detection typically performs i) encrypt the data. ii) Detecting malware with antivirus scanning the host. iii) Enforcing policies to limiting the transfer of sensitive data [10].

In host-based data leak detection approach, the data owner computes a specialized set of digests from the sensitive data and exposes only a small amount to Data Leak Detection (DLD) providers. The DLD provider computes the fingerprint from network traffic and identifies the potential leaks. The collection of leakage comprise of real leaks and inconsistent data [1], [15]. The data owner perform post-processes to determine whether there is any real leak by sending possible leak to the DLD provider. The data owner uses Robin fingerprint algorithm and a slipping window to generate the one way calculation through the fast polynomial modulus operation [10], [13].

To discover the data leak and accomplish the privacy, data owner generates the set of particular digests called fuzzy fingerprints id. Fuzzy fingerprint is used to hide sensitive data in the crowd or network traffic [10]. The DLD provider performs review on network traffic but the provider may attempt to learn the information about sensitive data. The DLD provider detects leaks by range-based comparison rather than an exact match. The DLD provider is a semi-honest antagonist [1].

The reminder of this paper is organized as follows. Section II, describes the Related Works. Section III, describes the Proposed Work. Section IV, describes the Experimental Evaluation and Results. Section V summarizes the Conclusion and Future Enhancement.

II. RELATED WORK

Ameya Bhorkar, Tejas Bagade, Pratik Patil and Sumit Somani (2015) introduce Decoy Information Technique. A malicious intruder can steal secret data of the cloud user in spite of provider taking forethought steps i) Do not allow the physical access, ii) Zero tolerance policy for intruders that access the information storage and iii) logging all access to the service and use internal inspection to find the malicious intruder [7]. Decoy is fake data which can be generated on requirement and serve as a mean of detecting unauthorized access of data. This technology is combined with user behavior to secure user information in cloud. Based on user activity clouds server send the original or decoy file to the user.

Amod Panchal, Grinal Tuscano, Humairah Kotadiya, Rollan Fernandes and Vikrant Bhat introduce Fake object, are basically changed in the original data which evidently improves the chance to find guilty agents [2]. The company may be able to add fake objects to the distributed data in order to improve distributor effectiveness in detecting guilty agents. The idea of disturbing or changing data to detect a leak is not a new technology. It can identify unauthorized use of data. The main disadvantage is it leads to modification of original data.

Brintha Rajakumari.S, Mohamed Badruddin .M and Qasim Uddin introduces secure login. Data distribution can be manipulated for better data mining to gain better conclusion and defend [4]. Data integration means that no sensitive data can be disclosed during data mining. Secure Hash Algorithm (SHA) provides more security and privacy data mining model. It addresses the problem of relinquishing private data. Different datasets composed of same set of user is held by the two parties. In data storage and encryption, the cloud server will partitioning the data into many portions when once data was stored into the web server; and store all the data in the separate server. Data and source code are stored in the data server. Admin is a person, who integrates the data in the web server.

Danfeng Yao,Elisa Bertino and Xiaokui Shu introduce fuzzy fingerprint data leak detection method. This approach focus on analyzing the unscripted outbound network traffic [1], [11] by i) deep packet inspection and ii) information theoretic analysis. In deep packet inspection, every packet is examined to identify the occurrence of sensitive data. This is also referred to as fuzzy fingerprint data leak detection model. The main characteristic of which is that the detection does not demand the data owner to reveal the content of stored sensitive data. Fuzzy fingerprint algorithm can be used to detect unexpected data leaks due to human error and / or application flaws and also this algorithm is used to hide sensitive data in a crowd.

Janga Ajay Kumar and K. Rajani Devi introduce guilty agent model. Data leak may find an unauthorized place in the establishment. It is either monitored or sometimes not monitored by the data owner. Data watcher model is used to identify data in leaks and guilt model capture leakage where agents can conspire and identify phony tuples. Guilt model is used to enhance the possibility of identifying guilty third parties [10], [13], [16]. It adds phony object to the distributed site is acting as a type of watermark for the entire set, without altering any particular member. If an agent gave one or more phony object that was leaked, then the distributor can be confident that the agent was guilty.

Chandni Bhatt and Richa Sharma introduce Watermarking Technique. Data leak detection is covered by water marking, e.g., a unique code is enclosed in each distributed copy. If that copy is later discovered in an unauthorized individual, the leaker can be identified [14], [15]. Watermarks can be very useful but involve some alteration of the original data. Example a company may have partnerships with other parties that require sharing customer's data. Another company may outsource its data processing, so data must be given to various other companies.

III. PROPOSED WORK

The system proposes a privacy-preserving data-leak detection model for preventing accidental data leak in network traffic. The DLD provider may findout sensitive information from the network traffic, which is inevitable for all deep packet inspection approaches.The proposed system uses the Secure Hash Algorithm (SHA) to generate short and hard-to-reverse digests through the fast polynomial modulus procedure. By using these techniques, an Internet service provider (ISP) can perform detection on its customers' traffic securely and provide data-leak detection as an add-on service to its customers.

However, privacy is a major barrier for recognizing outsourced data-leak detection. A conventional solution involves the data owner to reveal its sensitive data to the DLD provider. Existing work on cryptography-based multipartite calculation is not efficient enough for practical data leak inspection in this setting. Preventing sensitive data from being compromised is an important and practical research problem. The algorithms in this project achieve exact results by dismissing fields that are repeated or constrained by the protocol. The exposure of sensitive data in storage and transmission poses a serious threat to organizational and personal security.

Data leak detection aims at scanning content of exposed sensitive data. Because of the large content and data volume, such a screening algorithm needs to be scalable for a timely detection. Using special digests, the discloser of sensitive data is kept to a minimum during the detection. It presents a novel fuzzy fingerprint method for detecting accidental data leak in network traffic. This strong privacy guarantee yields a powerful application of fuzzy fingerprint method in the cloud computing environment, where the cloud provider can perform data-leak detection as an add-on service to its clients. The privacy model is useful beyond the specific fuzzy fingerprint problem studied. The detection is based on the fast set-intersection operation between the set of fingerprints generated from the payload of intercepted traffic (done by the DLD provider) and the set of fingerprints generated from the sensitive data.

This model supports detection procedure delegation and ISPs can provide data-leak detection as an add-on service to their customers using this model. And design, implement, and evaluate an efficient technique, fuzzy fingerprint, for privacy-preserving data-leak detection. Fuzzy fingerprints are especially sensitive data digests prepared by the data owner for release to the DLD provider.

A. Authentication Process

The system enables the data owner to securely delegate the content-inspection task to DLD providers without exposing the sensitive data. 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. 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. Sensitive data are sent by a legitimate user intended for legitimate purposes. The data owner is aware of legitimate data transfers and permits such transfers. So the data owner can tell whether a piece of sensitive data in the network traffic is a leak using legitimate data transfer policies.

B. Fuzzy Fingerprint Identification

To achieve the privacy goal, the data owner generates a special type of digest. The digests are called fuzzy fingerprints. The fuzzy fingerprint is to hide the true sensitive data in a crowd. It prevents the DLD provider from learning its exact value. The data owner chooses four public parameters. Data owner computes the set of all Rabin fingerprints of the piece of sensitive data. The data owner transforms each fingerprint into a fuzzy fingerprint.

C. Hashing Process

The system performs extensive experimental evaluation on both fingerprint filter and bloom filter with SHA algorithm and compares the runtime of Bloom filter provided by standard Pybloom from SHA-1 and that of fingerprint filter with Rabin fingerprint. SHA-1 produces a 160-bit (20-byte) hash value known as a message digest. A SHA-1 hash value is typically rendered as a hexadecimal number, 40 digits long.

D. Identifying Data Leak Detection through monitoring

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. 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.

E. Finding Shortest Path

By using SHA algorithm and DLD, the system can find the shortest path. The shortest path problem is the problem of finding a path between two vertices in a graph such that the sum of the weights of its constituent edges is minimized. The problem of finding the shortest path between two intersections on a road map may be modeled by a special case of the shortest path problem in graphs.

IV. EXPERIMENTAL ANALYSIS OF HOST-BASED DATA LEAK DETECTION AND RESULTS

Data owner processes the file and makes the fuzzy fingerprint then send the data to the Data leak detection (DLD) provider. DLD provider receives the file from the data owner and generates the hash key. DLD provider then compares the hash key which is generated by the data owner and DLD provider. If the matching key is differs, it information to the data owner that the data is leaked. Then DLD provider sends and monitors the file to the receiver through the host (shortest path). If any traffic occurs in transmission, the DLD provider sends the alert message to the data owner.

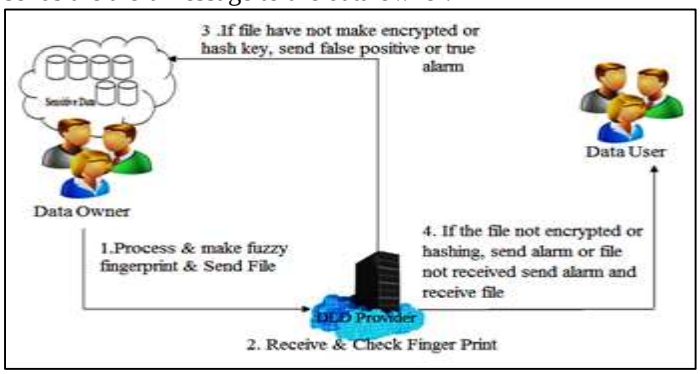

Fig. 1: Architecture of Data Leak Detection

Figure1 illustrates Architecture of the DLD model. In DLD model each and every data owner has its individual username and password. If the data owner wants to transfer the data to the receiver, data owner must perform the preprocessing operation to remove the noise and make the data as consistent. Fuzzy fingerprint is generated to hide the sensitive data in the crowd (network traffic). SHA algorithm is used to provide the hash value of the document. The datum is transferred through the DLD provider to identify potential data leaks. DLD provider again generates the hash value of the data and compared with the original hash value to identify leaks. Data owner computes the data leak which is sent by the DLD provider by comparing the fuzzy value with the original fuzzy fingerprint set. To overcome the traffic shortest path is used to transfer data from the data owner to the receiver with minimum traffic.

Data owner uses DES (Data Encryption Standard) algorithm to encrypt the sensitive data to maintain the privacy. The data owner generates a special type of digest. The digests are called fuzzy fingerprints. The fuzzy fingerprint is to hide the true sensitive data in a crowd. It prevents the DLD provider from learning its exact value. It used SHA to generate the hashing value of the sensitive data which are transferred.

A. Data Encryption Standard Algorithm

Data Encryption Standard (DES) is a block cipher cryptographic key. Algorithm is applied to a block of data simultaneously rather than one bit at a time. An algorithm that takes a fixed-length string of plaintext bits and translates it through a series of operation into another ciphertext bit string of same length as plaintext. Each block is deciphered using the secret key into a 64-bit ciphertext by means of permutation and substitution. It involves 16 rounds and can run in 4 different modes. DES uses a 64-bit key but 8 of those bits are used for parity check, so efficiently limiting the key into 56 bits (i.e.) 2^56 or 72,057,594,037,927,936 attempts to find the correct key.

It performs permutation on the 64-bit key of the permuted block is bit 57 of the original key. It split the permuted key into two halves. The first 28 bits are called c[0] and the last 28 bits are called d[0]. Calculate the 16 sub keys, starts with i=1. It performs one or two circular left shifts on both c[i-1] and d[i-1] to get c[i] and d[i] respectively. Then it split the block into two halves. The first 32-bits is called L[0] and the last 32-bits is called R[0]. It expands the 32-bit R[i-1] into 48-bits according to the bit selection function. DES performs XOR operation until 8 blocks have been replaced. Finally permute the concatenation of blocks.

B. Secure Hash Algorithm

SHA to generate short and hard-to-reverse digests through the fast polynomial modulus operation. SHA is a cryptographic hash functions designed by the SHA produce a 160-bit (20-byte) hash value known as a message digest. A SHA hash value is typically rendered as a hexadecimal number, 40 digits long. It runs in polynomial time if the number of operations can be bounded by a polynomial in the size of input when data items are encoded in binary. A hash function is simply an algorithm that takes a string of any length and decreases it to a unique fixed length string.

It picks a string followed by internalizing some variable and breaks the string into a character. Then it converts a character into ANCII code and covert ANCII into binary and adds 1 to the end. Append original message length by adding 0's to the end. Chunk the break first and break into words i.e. break each chunk up into sixteen 32-bit words. Then it performs XOR operation by internalizing some variables. Finally the variables are converted into base 16 (hexadecimal) numbers and join together.

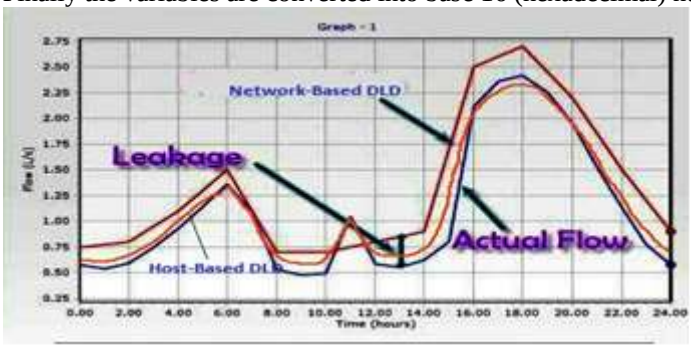

Fig. 2:Comparison between network-based DLD and Host-based DLD

Figure 2 illustrate the comparison between network-based DLD and Host-based Data Leak Detection. This secrecy preserving method provides better data leak prevention than network based DLD. It is inspecting the data's through preprocessing. Then data's is encrypted using DES algorithm and generating fuzzy fingerprint by using SHA. The Encryption and Hashing provide good strength to the sensitive data from data leak. That why host based data leak detection provide better performance than the network-based data leak detection. Finding the shortest path is another important metric to enhance the performance thereby reducing the network traffic.

V. CONCLUSION AND FUTURE WORK

The exposure of sensitive data in storage and transmission poses a serious threat to organizational and personal security. . Data leak detection aims at scanning content of revealed sensitive data. Preventing sensitive data from being compromised is an

important and practical research problem. The algorithms in this project achieve precise results by discounting fields that are repeated or constrained by the protocol. The system implemented, and evaluated a new privacy-preserving data-leak detection system called host-based DLD that enables the data owner to safely deploy locally, or to delegate the traffic-inspection task to DLD providers without exposing the sensitive data. The main feature is that the detection does not expose the content of the sensitive data and enables the data owner to safely assign the detection. Rather, only a small amount of specialized digests is needed. This host-based DLD technique may improve the high accuracy performance, good privacy and guarantee to reduce the Network Traffic.

REFERENCES

[1] Danfeng Yao,Elisa Bertino and Xiaokui Shu, "Privacy-Preserving Detection of Sensitive Data Exposure" IEEE Transaction on Info. Forensic and Security, Volume: 10, No. 5, May 2015.
[2] Francis Thivya A. and Tharcis P., "A Survey on Secrecy Preserving Discovery of Subtle Statistic Contents", American Eurasian Network for Scientific Information Publisher –Journal of Applied Sciences Research, pp. 14-19, Nov.2015.
[3] Amod Panchal, Grinal Tuscano, Humairah Kotadiya, , Rollan Fernandes and Vikrant Bhat , " A Survey On Data Leakage Detection", Int. Journal of Engineering Research and Applications, Vol. 5, Issue 4, (Part -6), pp.153-158, April 2015.
[4] Jung.J, Sheth.A, Greenstein.B, Wetherall.D, Maganis.G and Kohno.T,"Privacy Oracle: A system for finding application leaks with black box differential testing", in proc. 15th ACM Conf. Comput. Comm. Secur., 2008, pp.279-288.
[5] Brintha Rajakumari.S, Mohamed Badruddin .M and Qasim Uddin," Secure Login of Statistical Data with Two Parties" International Journal on Recent and Innovation Trends in Computing and Comm., Volume: 3, Issue: 3, March 2015.
[6] Butt. A.R, Liu.F, Shu.X, and Yao.D, "Privacy-preserving scanning of big content for sensitive data exposure with MapReduce", in Proc. ACM CODASPY, March 2015.
[7] Emiliano De Cristofaro, Gene Tsudik and Yanbin Lu, "Efficient Techniques for Privacy-Preserving Sharing of Sensitive Information", IEEE Transaction on Information Forensic and Security, March 2015.
[8] Ameya Bhorkar, Tejas Bagade, Pratik Patil and Sumit Somani," Preclusion of Insider Data Theft Attacks in the Cloud "International Journal of Advanced Research in Computer and Communication Engineering, Volume: 4, Issue: 1, Jan. 2015.
[9] Supriya Singh, "Data Leakage Detection Using RSA Algorithm ", International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume: 2, Issue: 5, May 2013.
[10] Cova. M, Kapravelos. A, Kruegel. C, Shoshitaishvili.Y and Vigna. G, "Revolver: An automated approach to the detection of evasive web-based malware", inProc. 22nd USENIX Secur. Symp. , pp. 637–652, January 2013.
[11] Janga Ajay Kumar and Rajani Devi .K," An Efficient and Robust Model for Data Leakage Detection System", Journal of Global Research in CS, Volume: 3, No.6, June 2012.
[12] Shu .X and Yao.D, "Data leak detection as a service", in Proc. 8th Int. Conf. Secur. Privacy Commun. Netw, pp. 222–240.
[13] Caesar. J and Croft. M, "Towards practical avoidance of information leakage in enterprise networks", in Proc. 6th USENIX Conf. Hot Topics Secur. , p. 7, June 2011.
[14] Archana Vaidya, Kiran More , Nivedita Pandey , Prakash Lahange and Shefali Kachroo, "Data Leakage Detection", International Journal of Advances in Engineering & Technology, Vol. 3, Issue 1, pp. 315-321, March 2012.