ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
IJARSMT
Volume 1, Issue 6, Nonember 2015

A Review on Fuzzy Fingerprints Based Privacy Preserving Approach

Minal Loharkar[1], Dr.Shyamrao Gumaste[2]

PG Student, Dept. Of CSE, R.H.Sapat College of Engineering, Nashik, Maharashtra, India[1]

Associate professor, Dept. Of CSE, R.H.Sapat College of Engineering, Nashik, Maharashtra, India[2]

ABSTRACT— Now-a-days there are many data-leaks. The Statistics from security firms, research institutions and government organizations shows many data-leaks. Human mistakes are one of the main causes of data loss in various data-leak cases. There exist many solutions detecting unintentional, sensitive data leaks caused by human mistakes and to provide alerts against them to the organization. Here privacy preserving data-leak detection (DLD) solution is used to detect the data loss. Two techniques are used Rabin fingerprint and sliding window. The advantage of proposed method is that, it enables the data owner how to protect the data from being leaked. DLD is offered to the customers with some Internet service providers that can offer service with strong privacy guarantees. Thus evaluation results show that this method can support accurate detection of leaks in various data-leak cases with very small number of false alarms.

KEYWORDS- Data leak, network security, privacy, collection intersection.

I. INTRODUCTION

Now-a-days the number of data leaks has been increased. The leaks are occurring due to human mistakes. To detect & prevent the data leaks there should be some techniques available. The techniques for different leaks are different. The behavior of data is needed in order to implement the suitable preventive measures .The losses are due to the confidential information, customer data, health records, inadvertent leaks, planned attacks. But the looses due to users mistakes are more. So to prevent the data leaks the DLD (data leak detection) solution is used where a special set of sensitive data is used for detection.

Now the sensitive data is the data in which IT systems usually saves data in a database user's personal information. The information such as passwords, credit card numbers, house address, telephone number, id number etc .When the system is not protected effectively from unauthorized access there is a high probability that a hacker might utilize the unprotected and steal that information. That vulnerability is Sensitive Data Exposure.

The propose method enables the data owner to safely delegate operation to a semi honest provider without providing the sensitive data to the provider. Here how Internet service providers can offer their customers DLD to improve the privacy guarantees is made.

ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
Volume 1, Issue 6, Nonember 2015

The main idea of privacy preserving is to relax the comparison criteria by introducing matching instance on the DLD provider's side without increasing the amount of false alarms, false alarms are the warning of something bad to happen but it does not happen to the data owner.

1) The data owner before disclosing them to the DLD provider, perturbs the sensitive-data fingerprints, and

2) Instead of the exact match the DLD provider detects leaking by a range-based comparison. During comparison range is pre-defined by the data owner and compares with the perturbation procedure.

A DLD provider is used to obtain digests of sensitive data from the data owner. The two methodologies are used by data owner Rabin fingerprint algorithm to generate short and hard to reverse digests through the fast polynomial modulus operation and sliding window [12].

II. LITERATURE SURVEY

The work done by J. Li, Q. Wang, C. Wang, N. Cao, K. Ren, and W. Lou[1] was it formalizes and while maintaining keyword privacy solves the problem of effective fuzzy keyword search over encrypted cloud data. Also it exploits edits distance to find keywords similarity by developing an advanced technique on constructing fuzzy keyword sets, which greatly reduces the representation overheads and storage. In this first a plaintext fuzzy keyword search is considered. The fuzzy search was started with an attention of plaintext searching in information retrieval community. It was done by allowing users to search without using try-and-see approach for getting the relevant information based on approximate string matching. But the construction to apply string matching algorithm was possible. Limitation was it suffers from dictionary and different attacks and fails to achieve privacy. The second consideration was searchable encryption. The study of cryptography is done in this stage. But the limitation was the above scheme support only exact keyword search and thus not suitable for cloud computing. The further propose an systematic technique fuzzy keyword search scheme. Through rigorous security analysis, shows that proposed solution is secure and privacy-preserving, while correctly realizing the goal of fuzzy keyword search.

J. Croft and M. Caesar [2] proposed a black-box differencing. It was run on two logical copies of the network, one with private data cleaning, and compare outputs of the two to determine if and when private data is being leaked. To ensure outputs of the two copies match, construction of recent advances that enable computing systems to execute deterministically at scale and with low overheads is done. The approach could be a useful building block towards building general-purpose schemes that hold black-box differencing to mitigate leakage of private data. Threats of leaks of sensitive data are a growing threat to networks that store sensitive data, such as source code or customer information. To this end, this paper proposes a network-wide method of data confinement that detects information leaks by forking copies of processes consuming private data and removing the sensitive data from the input to the copy. There is introduction of the concept of a paired packet to allow both copies of the process to send data onto the network to allow the sharing of sensitive data within the confines of the network.

In this paper, S. Ananthi, M. Sadish Sendil, and S. Karthik[3] a search plan that provides both privacy protection and rank-ordered search capable for holding with less overhead has been proposed. Search indexes and documents are first encoded by the data owner and then stored onto the cloud server. Retrieval results on an encoded data and security analysis under different attack models show that data security can be preserved while retaining very good retrieval performance. The technique proposed enable systematic search directly in the encoded domain, without multiple rounds of communications between the user and the server. By analyzing the requirements of secure search sequence of

events, a secure indexing plan that makes use of inverted indexes of keywords is proposed. This plan achieves systematic data retrieval and is scalable for large files. Jointly utilize cryptography and search techniques to ensure that the encoded search indexes can preserve the search capability.A search scheme that provides both privacy secure and rank-ordered search capability with less expense has been proposed. The retrieval results on an encoded data and security analysis under different attack models show that data privacy can be preserved while maintaining very good retrieval performance.

X. Shu and D. Yao [4] proposed a network-based data-leak detection (DLD) solution that complements host-based methods. Network-based data-leak detection focuses on analyzing unencrypted outbound network traffic through i) deep packet inspection or ii) information theoretic analysis. For the deep packet inspection approach, a straightforward solution requires examine every packet for the occurrence of any of the sensitive data in the sensitive database. Such solutions generate awareness if the sensitive data is found in the outgoing traffic. However, this simple solution requires storing sensitive data in plaintext in the detection system. In this model, the data owner computes a special set of digests or fingerprints from the sensitive data, and then discloses only a small amount of digest information to the DLD provider. These fingerprints have important properties, which prevent the provider from gaining knowledge of the sensitive data, while they enable accurate comparison and detection. A deep packet inspection is processed by DLD to identify whether these fingerprint patterns exist in the outbound traffic of data owner's organization or not.A novel privacy-preserving data-leak detection model and its fuzzy fingerprint realization is proposed. Using special digests, the exposure of the sensitive data is kept to a minimum during the detection.Experiments were conducted to validate the accuracy, privacy, and efficiency to solutions.

Y. Jang, S. P. Chung, B. D. Payne, and W. Lee[5], propose a way to capture richer semantics of the user's intent. The method is based on the observation that for most text-based applications, the user's intent will be displayed entirely on screen, as text, and the user will make updation if what is on screen is not what she wants. Based on this idea, implementation of prototype called Gyrus which enforces correct behavior of applications by capturing user intent is done. Using Gyrus, demonstration of how to stop destructive activities that manipulate the host machine to send destructive traffic, such as social network impersonation attacks, and online financial services fraud is done. The evaluation results demonstrated that Gyrus successfully stops modern malware, and analysis shows that it would be very challenging for future attacks to beat it. Finally, the performance analysis shows that Gyrus is a viable option for positioning on desktop computers with regular user interaction. Gyrus fills an important gap, enabling security actions that consider user aim in determining the legitimacy of network traffic.

B. Wang, S. Yu, W. Lou, and Y. T. Hou [6] proposed a scheme, achieves fuzzy matching through algorithmic design rather than ellobrating the index file. It also eliminates the need of a predefined dictionary and systematically supports multiple keyword fuzzy search without increasing the index complexity.This paper tackled the challenging multi-keyword fuzzy search problem over the encrypted data. There is proposal and integrated several innovative designs to solve the multiple keywords search and the fuzzy search problems simultaneously with high efficiency. The approach of leveraging LSH functions in the Bloom filter to construct the file index is novel and provides an systematic solution to the secure fuzzy search of multiple keywords. In addition, the Euclidean distance is adopted to capture the similarity between the keywords and the secure inner product computation is used to calculate the similarity score so as to enable result ranking. A basic scheme as well as an improved scheme in order to meet different security requirements is

proposed. Thorough theoretical security analysis and experimental evaluation using real-world dataset were carried out to demonstrate the suitability of proposed scheme for the practice usage.

In this paper, A. Kapravelos, Y. Shoshitaishvili, M. Cova, C. Kruegel, and G. Vigna,[7] present Revolver, a novel approach to automatically detect vague behavior in malicious JavaScript. Revolver uses systematic techniques to identify similarities between a large number of JavaScript programs,and to automatically interpret their differences to detect avoidance. More precisely, Revolver influence the observing that two scripts that are similar should be classified in the same way by web malware detectors differences in the classification may specify that one of the two scripts contains code designed to avoid a detector tool. Using large-scale experiments, shows that Revolver is effective at automatically detecting evade attempts in JavaScript, and its integration with existing web manalysis systems can support the continuous improvement of detection techniques. This paper introduced Storage Capsules, a new mechanism for securing files on a personal computer. Storage Capsules are similar to existing encrypted file containers, but protect sensitive data from malicious software during decryption and editing. The Capsule system provides this protection by isolating the user's primary operating system in a virtual machine. The Capsule system turns off the primary OS's device output while it is accessing confidential files, and reverts its state to a snapshot taken prior to editing when it is finished. One major benefit of Storage Capsules is that they work with current applications running on commodity operating systems. Finally,evaluation is the overhead of Storage Capsules compared to both a native system and standard virtual machines.There is finding that transitions to and from secure mode were reasonably fast, taking 5 seconds and 20 seconds, respectively.

H. Yin, D. Song, M. Egele, C. Kruegel, and E. Kirda[8] propose a system, Panorama, to detect and analyze malware by capturing this fundamental trait. From experiments, Panorama successfully detected all the harmful samples and had very few false positives. Also further by using Google Desktop as a case study, shows the system can accurately capture its information access and processing nature, and can confirm that it does send back sensitive information to remote servers in certain settings.A system such as Panorama will offer indispensable assistance to code analysts and researchers by enabling them to quickly understand the nature and inner workings of an unknown sample.A Panorama can accurately capture its information access and processing behavior, and confirm that it does send back sensitive information to remote servers. Also believe that a system such as Panorama will offer indispensable assistance to harmful analysts and enable them to quickly understand the behavior and inner workings of malware.

J. Jung, A. Sheth, B. Greenstein, D. Wetherall, G. Maganis, and T. Kohno,[9] proposed privacy oracle.Privacy Oracle automatically detects leaks by looking for differences in the network traces produced by several test runs of an application. Limitations of privacy oracle are Encrypted connections. Without access to the implementation details and the source code of the target applications, it might be difficult to collect plaintext from an application that implements its own encoded scheme. Privacy Oracle must be provided a mechanism to inspect messages in plaintext, since properly encoded messages will always look different even if they are giving exactly the same information. Others are Message reordering. And Traffic randomization.Privacy Oracle, a system is presented that uncovers applications leaks of personal information in transmissions to faraway servers. The black-box differential fuzzy testing approach that Privacy Oracle takes, discovers leaks even when the structure of information being leaked was previously unknown, and does so without requiring a deep of the computing system under study; thus approach is broadly applicable to a myriad of device architectures and software systems.

ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
IJARSMT
Volume 1, Issue 6, Nonember 2015

In this paper, Danfeng Yao, Xiaokui Shu, *Member, IEEE*, and Elisa Bertino,[12] a privacy preserving data-leak detection (DLD) solution to solve the matter where a special set of quick to detect the data digests is used in detection is proposed. The advantage of method is that it enables the data owner to safely give the detection operation to a semihonest provider without revealing quick to detect data to the provider. And how Internet service providers can offer their customers DLD as an add-on service with strong privacy guarantees.

III. PROBLEM DEFINITION

Data-leak detection (DLD) is a privacy preserving technique used to solve the issue where a special set of sensitive data digests are used in detection. Propose method enables the data owner to safely delegate operation to a semihonest provider without providing the sensitive data to the provider. Here how Internet service providers can offer their customers DLD to improve the privacy guarantees is made.

IV. PROPOSED SOLUTION

A DLD (Data leak detection) technique is described in proposed solution. First what are Fuzzy fingerprints?

Fuzzy fingerprint is a technique to protect the data from DLD provider yet they do not cause additional false alarms for data owner as data owner can quickly distinguish between true and false leaks instance.

The fuzzy fingerprint includes:

Fuzzy length A fuzzy fingerprint f is given, fuzzy length *d* is the number of the least significant bits in fingerprint f that may be perturbed by the data owner, and *d* which is used to generate fingerprints is less than the degree of the polynomial.

Fuzzy set Given, a collection of fingerprints f and a fuzzy length *d*, the fuzzy set S (f,*d*) is the number of distinct collection of fingerprints whose values differ from f by at most $2^d - 1$.

4.1 Shingles

The shingle and fingerprint process is defined as follows.

The shingle(q-gram) is a fixed size sequence of contiguous bytes.Example the 3-gram shingle set of string abcdefgh consists of 6 elements {abc,bcd,cde,def,efg,fgh}.

4.2 Rabin Fingerprints

The Rabin fingerprint is as follows:

F (A) = A (t) mod P (t)

Where, A= ($a_1, a_2, \ldots a_m$) is a binary string

P is a reducible polynomial.

An example:

110101 mod 101 = 11

$$(X^5 + X^4 + X^2 + X^1) mod (X^2 + X^1) = (X + 1)$$

Advantage:

It is one-way and fast [11].Rabin fingerprints scheme is a method for implementing fingerprints using polynomials over a finite field, and can be implemented with fast XOR, shift, and table look-up operations.

ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
IJARSMT
Volume 1, Issue 6, Nonember 2015

The Rabin fingerprint algorithm has a unique min-wise independence property, which supports fast random fingerprints selection for partial fingerprints disclosure [10].

4.3 Algorithm

1) The Rabin-Karp string matching algorithm calculates a hash value for the pattern, as well as for each character of given text to be compared.

2) If the hash values for particular subsequence are unequal, the algorithm will calculate the hash values for next character.

3) If the hash values are equal it will do a brute force comparison between the pattern and character.

4) Therefore there is only one comparison per text subsequence and brute force only needed when hash values match.

4.4 Tables & Figures

The Table 1 is taken from https://www.sans.org/reading-room/whitepapers/awareness/data-leakage-threats mitigation-1931.It shows the type of information leaks and corresponding percentage. From Table 1 the leakage of data is more due to Customer data.

The Fig. 1 shows privacy-preserving data-leak detection model. In this the two important players are Data Owner and DLD (Data leak detection) Provider. The data-leak detection is a six step process:

The Data owner will preprocess and prepare fuzzy fingerprints.

Data owner will release the fingerprints and send it to DLD provider.

After that DLD provider will monitor outbound network traffic.

DLD provider will detect the fingerprints.

DLD provider will report all data-leaks alerts to the Data owner.

He/She will Post process and identify true leak instances.

Table 1: Survey of average data leak analysis

Type of Information leak	**Percentage**
Confidential information	15 %
Intellectual property	4 %
Customer data	73 %
Health records	8 %

ISSN (Online) : 2454-4159
International Journal of Advanced Research in Science Management and Technology
Volume 1, Issue 6, Nonember 2015

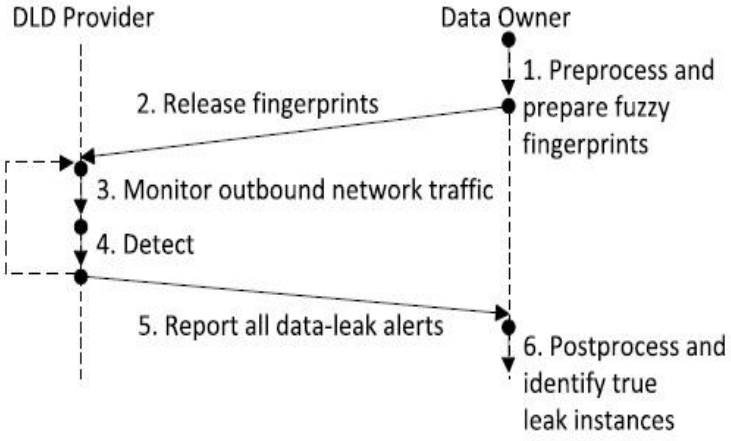

Fig. 1 Privacy-preserving Data-Leak Detection Model

V. CONCLUSION

Here a fuzzy fingerprint which is a privacy-preserving data-leak detection model and presentation of its realization is proposed. Using special digests, the exposure of sensitive data is kept to minimum during detection. The accuracy, privacy and efficiency are used to obtain solutions.

REFERENCES

[1] 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.

[2] J. Croft and M. Caesar, "Towards practical avoidance of information leakage in enterprise networks," in *Proc. 6th USENIX Conf. Hot Topics Secur. (HotSec)*, 2011, p. 7.

[3] 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.

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

[5] 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

[6] B. Wang, S. Yu, W. Lou, and Y. T. Hou, "Privacy-preserving multikeyword fuzzy search over encrypted data in the cloud," in *Proc. 33th IEEE Conf. Comput. Commun.*, Apr./May 2014, pp. 2112–2120.

[7] A. Kapravelos, Y. Shoshitaishvili, M. Cova, C. Kruegel, and G. Vigna, "Revolver: An automated approach to the detection of evasiveweb-based malware," in *Proc. 22nd USENIX Secur. Symp.*, 2013, pp. 637–652.

[8] H. Yin, D. Song, M. Egele, C. Kruegel, and E. Kirda, "Panorama: Capturing system-wide information flow for malware detection and analysis," in *Proc. 14th ACM Conf. Comput. Commun. Secur.*, 2007, pp. 116–127.

[9] J. Jung, A. Sheth, B. Greenstein, D. Wetherall, G. Maganis, and T. Kohno, "Privacy oracle: A system for finding application leaks with black box differential testing," in *Proc. 15th ACM Conf. Comput. Commun. Secur.*, 2008, pp. 279–288.

[10] M. O. Rabin, "Fingerprinting by random polynomials," Dept. Math., Hebrew Univ. Jerusalem, Jerusalem, Israel, Tech. Rep. TR-15-81, 1981.

[11] A. Z. Broder, "Some applications of Rabin's fingerprinting method," in *Sequences II*. New York, NY, USA: Springer-Verlag, 1993, pp. 143–152.

[12] Privacy-Preserving Detection of Sensitive Data Exposure Danfeng Yao, Xiaokui Shu, *Member, IEEE*, and Elisa Bertino, IEEE Transcations on information forensics and security, VOL. 10, NO. 5, May 2015.