IJCSE International Journal of Computer Sciences and Engineering Open Access

Survey Paper Volume-4, Issue-11 E-ISSN: 2347-2693

Survey on Data Leak Detection Algorithms

Thrupthy Mohanan A[1*], Jimy George[2] , Nisha S M[3] and Lemya Sainudeen[3]

e-mail: *thrupthymohan@gmail.com, jforjimsbabluoo7@gmail.com, nisharanjith1810@gmail.com, lemya.sain@gmail.com*

Available online at: www.ijcseonline.org

Received: Oct/26/2016 Revised: Nov/05/2016 Accepted: Nov/24/2016 Published: Nov/30/2016

***Abstract*—The data leak detection plays a major role in organizational industry. The data leak poses serious threat to online social Medias, sensitive datas and so on. We take two papers for this survey. Both of them belong to the area of information forensic and security. In first survey, paper develops a model [PPDLD] the model is based on fuzzy finger print method. The goal of this paper is to generate special type of digest is called fuzzy fingerprint. Rabin finger print algorithm is introduced here just for sampling. A filtering method is used during the digestion process. In second survey, paper deals the fast detection of transformed data leak. This paper suggests a preserving method based on alignment algorithm. The paper aim to detect long and inexact leak patterns from sensitive data and network. Detection is based on comparable sampling algorithm.**

Keywords-Leak Detection; Fuzzy finger print random generation; Rabin karp algorithm; Subsequence preserving; Alignment algorithm

I. Introduction

Information leaks are major problem of computer system. The data leak detection play a major role in Organizational industry .The data leak poses serious threat to online social medias, sensitive datas and so on. Moreover the data leak detection [DLD][1] is based on two approaches – that is host based and network based. Normally the network based data leak detection is used to provide more efficiency and one way computation over the network packet for sensitive information which analyze the content of unencrypted network packet.

The data leak detection normally performs packet inspection method and searches for position of leaked patterns .The detection require a plain text sensitive data. Many of the algorithms perform leak detection in network intrusion models .We know the encryption and decryption algorithms are well played in detection algorithm. The data leak detections are virtualized .When the network function is virtualized [VNF-Virtual Network Function][2] which integrates the cost to deploy and provide slice scheduling which yield resource and performance on node stream slicing[2].The node stream slicing consist control plane and data plane. Use a custom scheduling which generate aggregate throughput on node stream slicing.

Today, many of the social Medias are reported as leaked for example face book .Face book faces unauthorized access which occur on password en-decryption. Some of the leak detection algorithms [DDA] runs internally or externally. Its performance is based on hash value or threshold value .The hash value or threshold value yield effective efficiency and proper computation on band width. The band width gives high eccentric performance. When the requirements are unpredictable, it may cause sensitive information will threaten.

Our first survey paper focus the privacy preserving data leak detection [PPDLD][3,4]exposure on sensitive content. The PPDLD securely deliver content inspection task without exposing the data .They introduced a fuzzy finger print method [3]. For the secure delegation use data hiding in forensic way. Data hiding is a software development technique which is specifically used in object oriented concept to hide internal object details (data members).Sometimes the Data hiding [12] is byte oriented .Data hiding ensures exclusive data access to class members and protect object integrity by preventing unintended or intended changes. The hiding is based on stenography. Sometimes encoded decoded leak detection algorithms are proposed data hiding algorithms. This paper also introduce Rabin karp algorithm. Rabin karp [5, 6] is a type of leak algorithm. Rabin karp detects leak by tracking position of the data stored array.

Our second survey paper, “fast detection of transformed data leak” introduces the concept of preserving. The paper mainly aims to detection of long and inexact sensitive data pattern, and also detects leaked patterns from sensitive data and network. Generally using comparable algorithm for the detection of long and inexact leaked patterns. For the detection of leaked patterns use subsequence preserving sampling algorithm and alignment algorithm. The data driven semi global alignment algorithm [DDSGAA] [7] used in masquerade detection Using sampling method provide more space efficiency.

II. PROPOSED METHODS

To conduct comparison analysis study on leak detection algorithms and to infer the best approach in the field, two papers are considered under the content inspection task. But both paper taken two different methods for content inspections.

1. Fuzzy fingerprint method.
   Fuzzy finger print consist valuable data. The valuable data consist parameters. Using these parameters generate fuzzy finger print (32 bit) which is randomized. For the generation of random numbers use timestamp in sequence number.
2. Subsequence preserving method.
   In subsequence preserving, introduce a sampling [8] techniques. The sampling techniques accelerate pattern matching in intrusion detection systems. In sampling pattern recognition done according to regexes (regular expression).Mainly the sampling techniques introduced in string matching algorithms which yield extreme spatial efficiency.

III. RESULT AND DISCUSSION

Generate fuzzy finger print in privacy preserving on sensitive data exposure

In this work, they presented a novel data leak detection model named as privacy preserving data leak detection[PPDLD].The model proposed client server interactions. In this model data owner and dld provider[9] act as a client server.

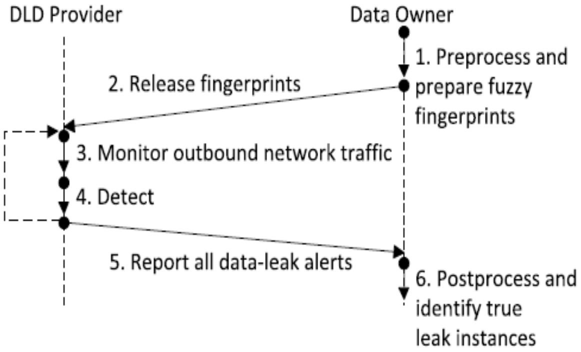
DLD Provider
Data Owner
1. Preprocess and prepare fuzzy fingerprints
2. Release fingerprints
3. Monitor outbound network traffic
4. Detect
5. Report all data-leak alerts
6. Postprocess and identify true leak instances

Fig1. Privacy preserving data leak detection model

Fig.1.Illustrate the privacy preserving data leak detection model. Working of this model is based on six step operation phase. 1) In preprocess stage, there is data owner's authentication done. After authentication the data owner generates fuzzy finger print. 2) Release fuzzy finger print to dld provide.3) Monitor the outbound network traffic.4) Detect. For the detection set sensitivity value. The dld provider do not exactly find out the value (1/k where k is represent a integer) When the value will greater than 1/k then leak detected on network traffic. Otherwise no leak found.5) Report all data leak determined by using alarm alert algorithm.6) Identify the true leak instance.

The fuzzy finger print consist valuable data. The valuable data consist some parameters (eg.transaction key) Using these parameters generate fuzzy fingerprint (32 bit, randomized).When 33 bit generate then the fingerprint is irreducible. For the generation of random numbers use timestamp in sequence number (sometimes used as a decoded information). Use java cryptography function for the generation of fuzzy finger print. A finger print filter used in detection stage only for filtering the leak. Bloom filter [BF] is used for space efficency. This [BF] is determined by in probabilistic way. The bloom filter determines whether the element is a given group or set. In filtering, apply a multiple hash values to elements or assign to sets and store values in bit vector .The combination of bloom filter and Rabin finger print is referred to as fingerprint filter. For the conversion of transaction key into secret key object use Hmac.MD5 [Message Digest Version 5] algorithm is used to create MAC object to generate the fuzzy fingerprint. Add caret (^) between input value as per documented by authorize net Sensitive data digest from byte array. A shingling method involved in between the conversion of transaction key to fuzzy finger print (secret key).

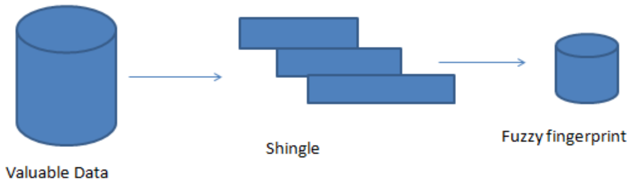

Fig2. Fuzzy finger print generation

Fig2.Illustrate the mode of fuzzy finger print generation. For shingling q-gram approach used. A sliding window is used to generate q-gram on an input binary string. After fuzzy finger print computed.

B. wang, S. *et.al* [10] proposed a paper on fuzzy keyword search. In fuzzy keyword search which search over the encrypted cloud data. In keyword search done interaction between user and owner. There is cloud server exist in between user and owner. The keyword search is one of the secured way of searching method in pattern recognition. Here the owner encrypted files to cloud server. The user access encrypted files from cloud server.

Rabin Karp algorithm:

*R*abin karp is one of the internally running algorithm. Which calculate the numerical value (hash value) for the data pattern (p). For each m character substring of text (t).Then it compare the hash values instead of comparing the actual symbols. If any match is found it compare the pattern with the substring. Otherwise it shifts to next substring of txt to compares the pattern s with the substring of t to compare with p. This algorithm is mainly applicable in detecting plagarism.The algorithm provides effective computation. The Rabin finger print algorithm is popular and effective rolling hash function.

Example 1:
Enter Text: 2359023141526739921
Enter Pattern: 314152
Results: Pattern found at position 7

Rabin karp algorithm is string searching algorithm that uses hashing to find a set of pattern string in a text. Consider n represent the length of text and pattern p of combined length m, its average and best case running time is O(n+m) but its worst-case time is O(nm).

Fast transformed detection of data leak.

X.shu and D.Yao.*et.al*.[11] are proposed this paper. The main concept behind the paper is to detection of long and inexact leak patterns. A comparable sampling algorithm introduced for the detection of leak patterns from sensitive data in content inspection task. The objectives of the paper are multithreading scalability and context aware selection. Multithreading scalability is serializable. The term high light to protection of single threaded object from overlapped client request. The overlapped request generate internal error .Use flip method (OR, AND, or DIFFERENCE) for the comparison, the comparison is instantaneous, which produce high detection accuracy and high throughput by tracking leaked patterns. as a result the amount of time proceeds.

The context aware selection is the part of the subsequence preserving algorithm. It compares its surrounding items according to selection function. The selection function services to current context. That means, the function is purely aligned to present content inspection task. The approaches are based on context aware packet. The context aware packet consist constraints and data.The packet describe the service request .Using this selection function the subsequence preserving algorithm become more deterministic and preserving.

Example 2:
Original lists:
5627983857432546397824366
5627983966432546395
Sampled sequence need to be aligned as:
--2---3-5---2---3-7-2-3—
--2---3-6---2---3—
Regular local alignment:
23523723
23623

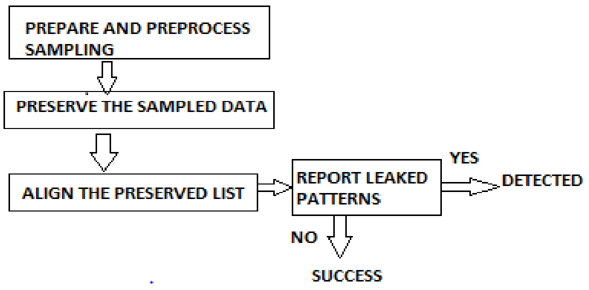

Fig3. *Fast transformed data leak approach*

As illustrated in Fig.3 leaked pattern detection from sensitive data or network. The figure introduce the concept of sampling, subsequence preserving and alignment algorithm [11]. Using sampling method is providing more extreme space efficency and accelerate pattern matching in leak detections. The sampling adopts the regular expression (regex) which defines a search patterns. After sampling use subsequence preserving sampling algorithm for preserving sampled data. For align the preserved list use alignment algorithm which is developed by using dynamic recurrence. The algorithm compare not only sampled data but also null regions (based on weight function) are also aligned. The results include match, mismatch and gap. There is a trace back in weight function which inferring matching based on null regions. When outcome show match no leaked pattern found, otherwise leaked pattern found.

Dominica ficara.*et.al*.[8] introduce a novel yet simple idea to accelerate DFA for security .Payload sampling used which skip a large portion of the text as a result processing less bytes .Use a run length encoding which encode the transition table. Sampling is focus to implementation of regex pattern matching. Regex buddy offers the way to getting fastness with regular expression.

Amol c. Devcate .*et.al* [7] introduce the concept data Driven semi global alignment algorithm. DDSGA is a derived version of semi global alignment algorithm. It is easily detect attack. For the masquerade detection use naive bayes therom .Naive bayes is a simple technique for constructing classifiers.DDSGA model that assign class labels to problem instances, represented as vector of feature values. Where the class labels are drawn from finet set.

Baye's theorem state that:
$P\ (x_i|y,x_1,....,x_{i-1},x_{i+1},......,x_n)=P(x_i|y)$ (1)
For all i,the relationship is simplified to
$P(y|x_1,....x_n)=P(y)¶_{i=1}^{n}\ P(x_i|y)\ /P(x_1,....,x_n)$ (2)

Three state of transition done diagonal, vertical and horizontal transition. These transitions provide more accuracy and efficency for semi global alignment algorithm, which exploit dynamic programming. Which initialize score matrix with order m+1 by n+1.This three transition are used to fill each cell in a transition matrix. DDSGA tries to avoid small mutation in user command which minimize overhead. The algorithm keeps consistency by providing different parameters to users. DDSGA model is security perspective based on sequence alignment. The main strategy behind the paper is to detect misalign area as masquerade by align the user informative summit sequence to previous one of the same user and labels.

Xiaoki shu,Jing Zhang *.et.al*[11] use subsequence preserving sampling algorithm and alignment algorithm for fast detection. The subsequence preserving sampling algorithm use a selection function. The resultant output of this algorithm is a sampled array. Existing method use set intersection here use set difference. In subsequence preserving, Generate a input sequence Initialize sliding window size .Set a collection difference for comparing the difference. Difference Taken based on selection function.

Example 3:
Consider Sliding window size |w|=6
N gram approach used
N=3
Input:1,5,1,9,8,5,3,2,4,8
The initial sliding window w= [1, 5,1, 9,8,5]
Collection m_c = {1,1,5}
Use Subsequence preserving sampling algorithm.
Output: 1-1----2--

Table1: Illustration of sampled procedure

Step	w	m_c	m_p	e_n	e_0	Sampled list
0	[1,5,1,9,8,5]	1,1,5	-	-	-	----------
1	[5,1,9,8,5,3]	1,3,5	1,1,5	3	1	1---------
2	[1,9,8,5,3,2]	1,2,3	1,3,5	2	5	1---------
3	[9,8,5,3,2,4]	2,3,4	1,2,3	4	1	1-1----2--
4	[8,5,3,2,4,8]	2,3,4	2,3,4	-	-	1-1----2--

Table 1: Shows four step of iteration done for getting sampled array.

For performing alignment algorithm need two sampled sequence. Set a threshold (T) value in sampled sequence. When the sequence get greater threshold compared to other is reported as leak. The alignment algorithm is more efficient and granular.

CONCLUSIONS

The first survey paper, proposed a privacy preserving data leak detection model .The model introduces a fuzzy finger print generation that is a special type of message digestion method using Hmac.MD5 algorithm on sensitive data exposure which provides more privacy and efficency. In second survey paper introduce fast transformed data leak detection which inspect content inspection technique for detecting leak of sensitive information in the content files or networks traffic. Detection is based on subsequence preserving and alignment algorithm. The result is suggest that the alignment method is better for detecting multiple common data leaks. Its prototype provide speed up and high scalability. For future work, we plan to explore fast transformed data leak detection on a host.

REFERENCES

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

[2] Robert riggo , Abbas Bradi , Davit Harutyunyan,Tinku Rasheed ,Member,IEEE,and toufik Ahmed "scheduling Wireless Virtual Networks Function" IEEE transactions on network and service management, VOL. 13, NO.2, MARCH 2016

[3] X. Shu, D. Yao, and E. Bertino, "Privacy-preserving detection of sensitive data exposure," IEEE Trans. Inf. Forensics Security, vol. 10, no. 5, pp. 1092–1103, May 2015.

[4] F. Liu, X. Shu, D. Yao, and A. R. Butt, "Privacy-preserving scanning of big content for sensitive data exposure with Map Reduce," in Proc. 5th ACM Conf. Data Appl. Secur. Privacy (CODASPY), 2015, pp. 195–206.

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

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

[7] H. A. Kholidy, F. Baiardi, and S. Hariri, "DDSGA: A data-driven semi-global alignment approach for detecting masquerade attacks," IEEE Trans. Dependable Secure Comput., vol. 12, no. 2, pp. 164–178, Mar./Apr. 2015.

[8] D. Ficara, G. Antichi, A. Di Pietro, S. Giordano, G. Procissi, and F. Vitucci, "Sampling techniques to accelerate pattern matching in network intrusion detection systems," in Proc. IEEE Int. Conf. Commun., May 2010, pp. 1–5.

[9] Chinar Bhandari1, Dr. Srinivas Narasim Kini "A Survey Paper on Data Leak Detection using Semi Honest Provider Framework" International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Index Copernicus Value (2013): 6.14 Impact Factor (2014): 5.611

[10] 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 IEEEConf. Comput. Commun., Mar. 2010, pp. 1–5.

[11] Xiaokui Shu, Jing Zhang, Danfeng (Daphne) Yao, and Wu-Chun Feng, Senior Member, IEEE "Fast Detection of Transformed Data Leaks" IEEE transactions on information forensics and security, VOL. 11, NO. 3, MARCH 2016

[12] Ankit Tale, Mayuresh Gunjal,B.A,Ahire "Data Leak Detection Using Information Hiding Techniques" IJCSE, VOL. 2, NO. 3, MARCH 2014

Authors Profile

***Mrs. Thrupthy Mohanan** A* pursed Bachelor of Computer Science and Engineering from University of Calicut, Kerala in 2012 and she is pursuing Master of Computer Science from A.P.J.Abdul Kalam

Technological University. Her main research work focuses on networking.

Mr. Jimy George pursuing a challenging role as a managerial teacher implementing key skills and knowledge towards the growth of the organization. Joined and started career as Assistant Professor in Department of Computer science and Engineering(RCET) from 2012 and pursued M.E in Computer and Communication from P.S.N.A.C.E.T, Dindigul in 2012 and Pursued B.E in Computer Science and Engineering from G.C.T, Salem in 2010.His main research work focus on Image processing, wireless sensor & ad-hoc networks, multimedia.

Mrs. Nisha S M pursed Bachelor of Computer Science and Engineering from University of Kerala, Kerala in 2012 and she is pursuing Master of Computer Science from A.P.J.Abdul Kalam Technological University. Her main research work focuses on Image processing.

Mrs. Lemya Sainudeen pursed Bachelor of Computer Science and Engineering from University of Calicut, Kerala in 2014 and she is pursuing Master of Computer Science from A.P.J.Abdul Kalam Technological University. Her main research work focuses on Networking.