View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Institute of Advanced Engineering and Science

International Journal of Electrical and Computer Engineering (IJECE)
Vol. 8, No. 4, August 2018, pp. 2247~2253
ISSN: 2088-8708, DOI: 10.11591/ijece.v8i4.pp2247-2253

A Survey: Data Leakage Detection Techniques

K. S. Wagh
Department of Computer Engineering, All India Shri Shivaji Memorial Society's Institute of Information Technology, Savitribia Phule Pune University, Pune, India

Article Info

Article history:

Received Sep 26, 2017
Revised Feb 19, 2018
Accepted Apr 13, 2018

Keyword:

Content inspection
Data leakage detection
Dynamic programming
Intrusion detection system
Sampling algorithm

ABSTRACT

Data is an important property of various organizations and it is intellectual property of organization. Every organization includes sensitive data as customer information, financial data, data of patient, personal credit card data and other information based on the kinds of management, institute or industry. For the areas like this, leakage of information is the crucial problem that the organization has to face, that poses high cost if information leakage is done. All the more definitely, information leakage is characterize as the intentional exposure of individual or any sort of information to unapproved outsiders. When the important information is goes to unapproved hands or moves towards unauthorized destination. This will prompts the direct and indirect loss of particular industry in terms of cost and time. The information leakage is outcomes in vulnerability or its modification. So information can be protected by the outsider leakages. To solve this issue there must be an efficient and effective system to avoid and protect authorized information. From not so long many methods have been implemented to solve same type of problems that are analyzed here in this survey. This paper analyzes little latest techniques and proposed novel Sampling algorithm based data leakage detection techniques.

Copyright © 2018 Institute of Advanced Engineering and Science.
All rights reserved.

Corresponding Author:

K. S. Wagh,
Department of Computer Engineering,
All India Shri Shivaji Memorial Society's Institute of Information Technology,
Savitribia Phule Pune University, Pune, India.
Email: waghks@gmail.com

1. INTRODUCTION

Data leakage in nothing but a getting access of a important data of a person or of an organization by unauthorized user. Important data of organization can have information customer, business plans, financial condition, information of patients, credit-card data of employs and so on depending on the business or a person or industry. But in various cases owner must share its data with its employees like when an employees are working for home using his own device or with the customers etc.

This will raise a possibility of important data falling into wrong hands. Which may be the result of some accident or a mistake by the person may be from the one who is working in the organization or by the outsiders such as hackers, can affect the organization. The organization can suffer from major damage due to data leak. Present systems for data leak detection are depending on set intersection. Set intersection is done on two sets of *n*-grams, one from the content and one from sensitive data. The set intersection gives the number of sensitive *n*-grams appearing in the content.

In such a way data leakage become the huge issue for various individual user, industries and institutes. A huge problem of the honesty of the users of those organization/systems is raised. It will tough for the person to find out the data leak in the various users. It also can create ethical problems in working organization. The potential harm and antagonistic results of an information leak event can be requested into the two classes: direct and indirect loss. Direct loss link to unmistakable harm that is definitely not hard to

measure and assess quantitatively. Indirect loss is much harder to measure and has a substantially more extensive effect regarding cost and time. Direct loss alludes to breaking the guidelines can come about up loss of future sales, expenses of examination and medicinal/rebuilding charges. Indirect loss can be result in less share-cost as a consequence of the negative publicity; can influence the reputation of organization or Intellectual Property to competitors.

Novel solution is for searching the transformed leakage in information using a sequence alignment algorithm, which is executed on the sampled delegate data sequence as well as the sampled substances. The alignment process scores are showing the measure of delicate information having in the substance. Our solution which is alignment-based, count the order of n-grams. It likewise takes care of arbitrary varieties of patterns without an unequivocal particular of all conceivable variety patterns.

2. LITERATURE REVIEW

In this paper [1], author makes use of sequence alignment method for searching complex data-leakage patterns. This algorithm is engaged for recognizing long as well as important data patterns. This identification is paired with a sampling algorithm that allows one to look at the similitude of two independently tested successions. This structure accomplishes great discovery exactness in perceiving transformed leakage.

Paper [2] author, implemented two algorithms for searching and transformed leakage information. This framework fulfills high recognition exactness and finds transformed leakage appeared differently in relation to the cutting edge inspection systems. They parallelize their design on graphics preparing unit as well as exhibit the solid scalability of their detection solution needed by a sizable association. In paper [3] authors have designs fuzzy fingerprint, which is a privacy-preserving data-leak detection system also provides its realization. By making use of special digests, the exposure of the vital data is kept to very less while detection. Authors have conducted few tests to conform the accuracy, privacy, as well as efficiency of our solutions.

In paper [4] author developed the Aquifer security system that assigns host export limitations on all data taken as part of a user interface (UI) workflow. Key understanding was that when applications in modern working frameworks offer data, it is a piece of an enormous work process to play out a user task. Each application on the UI work process is a potential information owner, and in addition thus can add to the security limitations. The restrictions are held with data as it is composed to storage and propagated to future UI work forms that read it. In doing all things considered they engage applications to sensibly hold control of their data after it has been shared as a major aspect of the client's tasks.

In paper [5] authors present Attire: an app for computers as well as smart phones which shows the user with an avatar. Attire conveys real-time data exposure in a light weight and unobtrusive manner via updating to the avatar's clothing. In paper [6] authors given the Data-Driven Semi-Global Alignment (DDSGA), DDSGA method. From the point of security effectiveness, DDSGA increase the scoring systems by adopting distinct alignment parameters for every single user. Also, it endures few transformations in user command sequences by permitting few changes in the low-level representation of the commands functionality. It additionally adjusts to modification in the user conduct by upgrading the signature of a user as per its present behavior. To optimize the run time overhead, DDSGA reduce the alignment overhead as well as parallelizes the detection and also modify.

In paper [7] author, proposed novel method for getting richer semantics of the user's determined. The technique is depending on the observation which for most text-based applications, the user's determined are shown fully on screen, as text, as well as the user will do some modifications if what is on screen is not what he needed. Depending on this concept, development of prototype known as Gyrus that enforces right working of applications by taking user determinant is done. By making use of Gyrus, representation of stopping destructive activities which can modify the host system to forward destructive traffic, like social network impersonation attacks, as well as online financial services fraud is done. The evaluation outcome shows that Gyrus successfully prohibits modern malware, as well as study demonstrated that it would be very tough for future attacks to defeat it. At last, the performance analysis demonstrated that Gyrus is a countable option for positioning on standalone pc with continues user interaction. Gyrus fills an important gap, enabling security actions that taking user concentration in finding the legitimacy of network traffic.

In paper [8] authors implemented a domain-specific concurrency model that backs a large class of IDS analysis not depends on a particular detection technique. Implemented technique divides the stream of network events in subsets that the IDS will process not related, as well as, while making sure each subset has each event relevant to a detection case. Proposed partitioning method is based on the concept of detection scope, i.e., the less "slice" of traffic that a detector must study for performing its function. As this concept has

some common applicability, designed model will support simple, per-flow detection technique and more complex, high-level detectors.

Findings of author [9], the introduction of essential data is not basic because of information change in the content. Transformations (for example, insertion, and deletion) results in significantly unpredictable leakage patterns. Present automata-based string coordinating algorithms are illogical for finding transformed data leakage as a result of its formidable complexity nature while exhibiting the required consistent expressions. They create two novel algorithms for recognizing long and also wrong data leakage. Their framework achieves high detection precision in perceiving changed breaks contrasted and the best in class inspection techniques. They parallelize our design on graphics processing unit and in addition demonstrated the solid scalability of data leakage detection arrangement examining enormous information.

Paper [10] authors given that number of the apparent distance metrics utilized for computing behavioral similarity between network hosts fail to capture the semantic significance imbued by network protocols. Moreover, they also tend to neglect long-term temporal structure of the objects being counted. To consider the role of these semantic as well as temporal attributes, they create another behavioral distance metric for network hosts as well as compare its execution with a metric which disregards information like this. Specifically, they propose semantically important metrics for common data types found in network data, indicate how these metrics can be consolidated to treat network information as a unified metric space, as well as depict a temporal sequencing algorithm which captures long-term causal relationships.

Shoulin Yin *et al*. [11] introduced novel concept searchable asymmetric encryption, which is useful for security and search operations on encrypted data. It greatly enhances the information protection, and prevent the leakage of the user's search criteria-Search Pattern. In paper [12] authors describes the Kaman-Kerberos assistant mobile ad-hoc network (KAMAN) protocol to avoid users information leak in cloud environment for virtual side channel attack. Moez Altayeb *et al*. [13] described the concepts of radiation leaks and data in wireless sensor network. To locate leakage station and control the stations power consumption by sending a special command to it from server node.

Table 1 shows the various authors papers details with method used, advantages and disadvantages.

Table 1. Various Authors Papers Details with Method Used, Advantages and Disadvantages

Sr. No.	Title	Paper Details	Method Used	Advantages	Disadvantages
1.	Fast Detection of Transformed Data Leaks	Utilize sequence alignment techniques for detecting complex data-leak patterns.	Comparable sampling algorithm alignment as well as sampling-oblivious algorithm.	This prototype provides substantial speedup and indicates high scalability of the design.	data-movement tracking approached is not used.
2.	Rapid screening of transformed data leaks with efficient algorithms and parallel computing	Design two novel algorithms for searching long as well as transformed information leaks.	Sequence alignment algorithm	This technique has high level of precision in finding transformed information leaks compared with the state-of-the-art set intersection technique.	Time Consuming Process.
3.	Privacy-Preserving Detection of Sensitive Data Exposure	Provides a privacy preserving data-leak detection (DLD) answer for the problem in which a special set of important information digests is utilized in detection.	MapReduce algorithm	Capability to arbitrarily scale as well as use of public resources for the process	System is not developed for intentional information exfiltration, which typically uses strong encryption
4.	Preventing accidental data disclosure in modern operating systems (2013)	Develops Aquifer as a policy system as well as system for avoiding accidental data disclosure in modern operating systems.	Aquifer as a policy system as well as framework For avoiding accidental data disclosure in modern operating systems	In Aquifer, application developers give secrecy restrictions which protect the entire user interface workflow during the user task.	Malicious applications are not taken into considerations are not taken into consideration
5.	Attire: Conveying information exposure through avatar apparel (2013)	Developed Attire: an app for computers as well as smart phones which displays the user with an avatar.	Attire: Mobile app	Attire passes real-time data exposure in a lightweight as well as unobtrusive manner via updating to the avatar's clothing.	This system can be modified to handle other sort of data with location, Such as views of photographs, `mentions' as well as `retweets' of

Sr. No.	Title	Paper Details	Method Used	Advantages	Disadvantages
					tweets, comments on status messages, etc.
6.	DDSGA: A data-driven semi-global alignment approach for detecting masquerade attacks	Given the Data-Driven Semi-Global Alignment, DDSGA method. From the security effectiveness view point, DDSGA upgrades the scoring systems by adopting distinct alignment parameters for every user.	DDSGA approach	DDSGA results improve both the hit ratio as well as false positive rates with an acceptable calculation overhead.	Need to detect masquerade attacks in cloud environment by improving this CIDS framework
7.	Gyrus: A framework for user-intent monitoring of text-based networked applications	Develop a way to break this cycle by making sure which a system's behavior matches the user's intent.	Gyrus framework	Gyrus successfully stops modern malware	The present design can be adapted to operate in a cloud computing model in which the remote host is an instance in an IaaS cloud. It must implement Gyrus on other platforms.
8.	Beyond pattern matching: A concurrency model for stateful deep packet inspection	Given a novel domain-specific concurrency model which solves this issue by having the notion of detection scope: a unit for partitioning network traffic like the traffic having in every resulting "slice" is independent for detection purposes	Intrusion detection systems (IDSs)	This technique correctly partitions existing sequential IDS studiesby having accuracy, while exploiting the network traffic's inherent concurrency potential for throughput upgrades.	It is complex process
9	Rapid and parallel content screening for detecting transformed data exposure.	Developed two novel algorithms for finding long as well as inexact information leaks.	Sequence comparison technique	This technique is can be used for big dataanalytics as well asfinding transformed sensitive information exposure.	Time consuming Process.
10	On measuring the similarity of network hosts: Pitfalls, new metrics, and empirical analyses	Proposed a behavioral distance metric for network hosts as well asanalyzed its performance to a metric whichneglectsdata like that.	Dynamic Time Warping (DTW)	Method gives consistent and useful characterizations of host behavior compared with the L1 metric.	This method not work on allow for localized reordering of points
11	Distributed Searchable Asymmetric Encryption	In this paper author discussed prevention of leakage of the user's search criteria-Search Pattern.	Polyfunctional concept of Searchable encryption.	To get higher efficiency and security in information retrieva	Complex search query and the leaking of search increases execution time.
12	KAMAN Protocol for Preventing Virtual Side Channel Attacks in Cloud Environment	In this paper author provides solution for cloud environment to avoid the leakage of user information.	Virtual side channel attack is described and how this attack provides the users credential to access user information.	Kaman- Kerberos assistant mobile ad-hoc network (KAMAN) protocol is used to avoid users informarmation leak.	Work is presented controlled environment, in real cloud environment, how provided solution is effective not discussed.
13	Wireless Sensor Network for Radiation Detection	In this paper, when wireless sensor network (WSN) send or recive data from other nodes. It reports radiation leaks, in addition to the data.	To locate leakage station and control the stations power consumption by sending a special command to it from server node.	GSM network is used to avoid radiation leak and data leak.	Effect of environmental paramers, while transmitting data in wirless sensor network.

3. RESEARCH METHOD

Four ways, we can secure organization data:

Identify Critical data: organization identify sensitive data of organization and use latest algorithm or software to protect critical data. Based on organization policies classify data, educate the users and provide sensitive data to employee based on access control mechanism.

Monitor Network Access: to monitor all network traffic and generate real time network access report. System find anomalous behavior pattern and auto notify this pattern to network administrator and block access control suspicious user.

Adaptive Security Mechanism: to provide adaptive to change security mechanism as per attacks for data leak.

Apply Security or Encryption Techniques: Apply excretion algorithm to critical sensitive information.

4. PROPOSED SYSTEM

The architecture and flow of the proposed system in Figure 1:

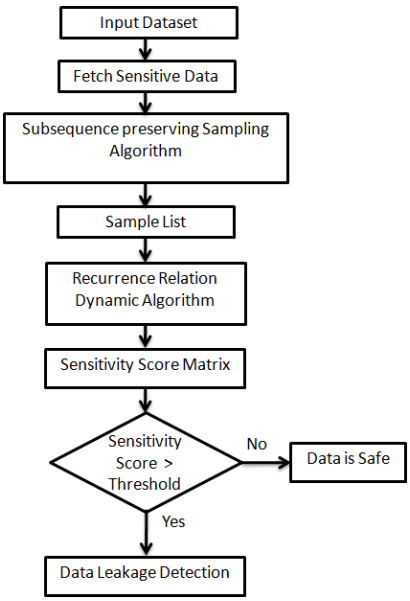

Figure 1. The architecture and flow of the proposed system

In the proposed system initially user browses the input dataset and fetches the sensitive data from input dataset. For generating the sample list, subsequence preserving sampling algorithm is used to generate sampling algorithm.

If given string is p and its substring is q, this is denoted by p ⊆ q), then p 0 is also a substring of q 0 (p 0 ⊆ q 0), where p 0 is a sensitive data of p sample and sampled sensitive data sample of p, and q 0 is a sensitive sampled data of q.

After that matrix for sensitivity score is calculated by recurrence relation dynamic programming algorithm. After that system is compared threshold value with sensitivity score value if sensitivity score is greater than threshold value then data leak is detected, otherwise data is not leak.

5. RESULTS AND DISCUSSION

Sampling algorithm based on context-aware selection. Selection decision of sample list depends on selection function, how selection compare with surrounding data. Sampling algorithm gives results as a deterministic and preserving the subsequence.

Recurrence Relation in Dynamic Programming algorithms works on compact sample list. This algorithm is alignment-based to detect data leaks using order-aware comparison and provides high tolerance to pattern variations. Also supports partial leak detection with high data leak detection.

System is implemented in C++ and dataset is used Enron [14], that contains 150 users email data with full header and bodies. To find data leak detection rate and false positive rate, using Equations (1) and (2).

$$Detection\ rate = \frac{\text{TP}}{\text{TP+FN}} \quad (1)$$

$$False\ positive\ rate = \frac{FP}{FP+TP} \quad (2)$$

where TP is true positive, FP is false positive, FN is false negative and TN is true negative. Table 2 shows data leak and no data leak in terms of above equations.

Table 2. Data leak in Terms of TP, FP and no Data leak in Terms of FN, TN

	True Leak	No Leak
Data Leak Detected	TP	FP
No Data Leak Detected	FN	TN

6. CONCLUSION

In real world, the organizations are facing the problem of data leakage. The data may be seen in other laptops or websites. From this survey we conclude that the data leakage detection system is largely useful for protecting the illegal use of data of various industries. So there is need to develop a content inspection method which detect leaks of important data in the content of files or network traffic. Also proposed system is useful to detect modification in data. In future such systems are necessary to detect data leak of personal, finance transactions, online shopping, and social media and so on.

REFERENCES

[1] X. Shu, *et al.*, "Fast Detection of Transformed Data Leaks," in *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 3, pp. 528-542, 2016.

[2] X. Shu, *et al.*, "Privacy-preserving detection of sensitive data exposure," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 5, pp. 1092-1103, 2015.

[3] F. Liu, *et al.*, "Privacy-preserving scanning of big content for sensitive data exposure with MapReduce," in *Proc. 5th ACM Conf. Data Appl. Secur. Privacy (CODASPY)*, pp. 195-206, 2015.

[4] A. Nadkarni and W. Enck, "Preventing accidental data disclosure in modern operating systems," in *Proc. 20th ACM Conf. Comput. Commun. Secur*., pp. 1029-1042, 2013.

[5] R. Hoyle, *et al.*, "Attire: Conveying information exposure through avatar apparel," in *Proc. Conf. Comput. Supported Cooperat. Work Companion (CSCW)*, pp. 19-22, 2013.

[6] H. A. Kholidy, *et al.*, "DDSGA: A data-driven semi-global alignment approach for detecting masquerade attacks," *IEEE Trans. Dependable Secure Comput*., vol. 12, no. 2, pp. 164-178, 2015.

[7] Y. Jang, *et al.*, "Gyrus: A framework for user-intent monitoring of text-based networked applications," in *Proc. 23rd USENIX Secur. Symp*., pp. 79-93, 2014.

[8] L. D. Carli, *et al.*, "Beyond pattern matching: A concurrency model for stateful deep packet inspection," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur*., pp. 1378-1390, 2014.

[9] X. Shu, *et al.*, "Rapid and parallel content screening for detecting transformed data exposure," in *Proc. 3rd Int. Workshop Secur. Privacy Big Data (BigSecurity)*, pp. 191-196, 2015.

[10] S. E. Coull, *et al.*, "On measuring the similarity of network hosts: Pitfalls, new metrics, and empirical analyses," in *Proc. Netw. Distrib. Syst. Secur. Symp. (NDSS)*, pp. 1-16, 2011.

[11] S. Yin, *et al.*, "Distributed Searchable Asymmetric Encryption," in *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 4, no. 3, pp. 684-694, 2016.

[12] H. Kaur and M. Kaur, "KAMAN Protocol for Preventing Virtual Side Channel Attacks in Cloud Environment," in *TELKOMNIKA (Telecommunication Computing, Electronics and Control)*, vol.15, no. 1, pp. 184-190, 2015.

[13] M. Altayeb, *et al.*, "Wireless Sensor Network for Radiation Detection," *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, vol. 5, no. 1, pp. 37-43, 2017.

[14] C. Kalyan and K. Chandrasekaran, "Information leak detection in financial e-mails using mail pattern analysis under partial information," *Proc. 7th WSEAS Int. Conf. Appl. Informat. Commun. (AIC)*, vol. 7, pp. 104-109, 2007.

BIOGRAPHY OF AUTHOR

Kishor S. Wagh received the BE degree in Computer Engineering from North Maharashtra University, Jalgaon , Maharashtra, India, in 1996, ME in computer Engineering in 2006 from PICT, Savitribai Phule Pune University, Pune and PhD in Computer Science and Engineering from the Swami Ramanand Teerth Marathwada University, Nanded. He is Associate Professor with the department of Computer Engineering, All India Shri Shivaji Memorial Society's Institute of Technology, Pune. His research area includes Web Services, wired and wireless networks, could computing, Mobile computing and IoT. He has published several research papers in refereed journals and professional conference proceedings. He is member of CSI, IEEE and Life member of ISTE.