ISSN No. 0976-5697

Volume 8, No. 5, May-June 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

A Literature Survey on Data Leak Detection And Prevention Methods

Aiswarya Baby
P.G Scholar
Department of Computer Science & Engineering
FISAT, Mookkannoor, Kerala, India

Hema Krishnan
Assistant Professor
Department of Computer Science & Engineering
FISAT, Mookkannoor, Kerala, India

Abstract:Data leakage occur when sensitive data falls into unauthorized hands. Sensitive data means intellectual property, financial data, patient data, personal credit-card data, and other confidential information depending on the business and the industry. Data leakage is an important issue for companies as the number of incidents and the cost to those experiencing them continue to increase. Whether caused by malicious intent or an inadvertent mistake by an insider or outsider, exposure of confidential information can seriously hurt an organization. This paper focuses on several data leak detection and prevention methods.

Keywords:Data Leak Detection; Data Leak Prevention; Time stamp; Confidential;Watermarking; Security

1. INTRODUCTION

The growth of modern technologies imposes a special means of security mechanisms. Security of data has become a major concern now a days. In the age of the internet, protecting our data has become important as protecting our property. We need to protect both physical and digital information from destruction and unauthorized access. Information is the primary product in the world of E-Commerce. As technology improves and access to markets expand, the need to secure information also increases. And to ensure its confidentiality, integrity, and authentication to those who need it for making critical personal, business, or government decisions become more important.

Data leakage and data misuse are serious threats to organizations. It is more severe when this is carried out by insiders. Sometimes it is very difficult to detect insiders. Data leakage is defined as accidental or unintentional distribution of private or sensitive data to unauthorized party [1]. Data leakage poses a serious issue for companies as the number of incidents and the cost those experiencing them continue to rise. Data leakage is magnified by the fact that transmitted data are not regulated and monitored on their way to their destination.

Security should go with the organization. If security measures are not properly implemented it can kill the confidentiality of the company. It can remove trust the customers have on the company. If one person in the entire employee segment do not have the security discipline he can potentially be a security vulnerability. That makes the problem much more complex. Kevin Mitnick [2] a renowned security consultant states that “People are the weakest link. You can have the best technology, firewalls, intrusion-detection systems, bio-metric devices and somebody can call an unsuspecting employee. That's all she wrote, baby. They got everything."

Data leak generally occur in 3 places. Inside, data were leaked from a source residing within the organizations physical perimeter; data were leaked from an external source residing outside the organizations perimeter; and Third-party location, data were leaked from a trusted third-party location. To avoid sensitive data loss a multilevel security system is to bedeveloped to protect those valuable data. Data leak detection differs from the anti-virus scanning (e.g., scanning file systems for malware signatures) or the network intrusion detection systems (NIDS). A significant portion of the data leak incidents are due to human errors. In order to minimize the exposure of sensitive data and documents, an organization needs to prevent clear text sensitive data from appearing in the storage or communication.

The remainder of this paper is organized as follows: Related works in the literature of data leak detection methods and data leak prevention methods are analyzed in Section II. Finally, a brief conclusion is given in Section III.

2. RELATED WORKS

This section gives an analysis on the various works that have been proposed in the area of data leak detection and data leak prevention.

In [3]authors focus on privacy-preserving detection of sensitive data exposure. They presented a data-leak detection solution which can be outsourced and deployed in a semi-honest detection environment. The advantage of their method is that it enables the data owner to safely delegate the detection operation to a semi-honest provider without revealing the private data to the provider.

They used fuzzy fingerprint technique that enhances data privacy during data-leak detection operations. The data owner preprocess and prepares fuzzy fingerprints and release the fingerprints to DLD provider. The DLD provider computes fingerprints from the 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. He reports all data leak alerts to the data owner. Data owner then post-processes the potential leaks sent back by the DLD provider and decides whether there is any real data leak.

In [4] authors focus on inadvertent leak detection. Detecting the exposure of sensitive information is challenging due to data transformation in the content. Transformations (such as insertion and deletion) result in highly unpredictable leak patterns. In the data leak detection model, they analyze two types of sequences: sensitive data sequence and content sequence. Content sequence is the sequence to be examined for leaks. The content may be data extracted from file systems on personal computers, workstations or payloads extracted from supervised network channels. Sensitive data sequence contains the information (e.g., customer's records, proprietary documents) that needs to be protected and cannot be exposed to unauthorized parties. The sensitive data sequences are known to the analysis system. Here they utilized sequence alignment techniques for detecting complex data-leak patterns.

In [5] authors formalize problem of provably associating the guilty party to the leakages, and work on the data lineage methodologies to solve the problem of information leakage in various leakage scenarios. They define LIME, a generic data lineage framework for data flow across multiple entities in the malicious environment. Three characters are involved- owner, consumer, and auditor. Auditor determine a guilty party for any data leak, and define the exact properties for communication between these roles.

The key advantage of the model is that it enforces accountability by design. This helps to overcome the existing situation where most lineage mechanisms are applied only after a leakage has happened. They present an accountable data transfer protocol to transfer data between two entities. To deal with an untrusted sender and an untrusted receiver scenario associated with data transfer between two consumers, the protocols employ an interesting combination of the robust watermarking, oblivious transfer, and signature primitives. Cox algorithm is used for watermarking.

In [6] the authors study unobtrusive techniques for detecting leakage of a set of objects or records. They developed a model for finding the guilty of agents. They also present algorithms for sharing objects to agents, in a way that enhances the chances of identifying a leaker. Finally, they also considered the choice of adding fake objects to the distributed set. Such objects do not match to real entities but come into sight realistic to the agents. In a sense, here the fake objects act as a type of watermark for the entire set, without modifying any separate members. If an agent was given one or more fake objects that were leaked, then the distributor can be more assured that agent was guilty.

In [7] authors focus to data leakage prevention system with a time-stamp. In Data Leakage Prevention, the time stamp is very important for giving permission to access a particular data, as in a particular period of time the data is confidential after the time stamp the same data could be non-confidential. In time stamped based DLP two phases are there, Learning Phase and Detection Phase.

In learning phase collect confidential and non-confidential documents of an organization. Then create clusters using K-means with cosine similarity function. For each cluster identify the key terms based on their frequency. For each key term calculate the score and assign time stamp for a document based on deadlines of organization schedule. In the detection phase the tested document is compared with the confidential score and time stamp, if the time stamp of the tested document is greater than or equal to the time stamp then that document is treated as a confidential and it is blocked.

In [8] a new context-based model for accidental and intentional data leakage prevention is proposed. The context-based approach they proposed leverages the advantages of preventing data leakage by either looking for specific keywords and phrases or by using various statistical methods. Their new model consists of two phases: training and detection. During the training phase, they created clusters of documents. Then a graph representation of the confidential content of each cluster is generated. This representation consists of key terms and the context in which they need to appear in order to be considered confidential.

During the detection phase, document tested is assigned to several clusters. Its contents are then matched to each cluster's respective graph in an attempt to determine the confidentiality of the document. One of the main advantage of their method is It detects small sections of confidential information embedded in non-confidential documents. It generates a well-understood model that can be reviewed and even modified by its users.

In [9] authors aims to prevent the data leakage stemming from corporate email. When, an employee sends an email, which contains an attachment, from his corporate account to a recipient, the generated email is forwarded to the SMTP port which accepts outbound emails, on his system. SMTP proxy server can picks up the email and trigger the steganography scanner. Attachments are scanned and if they are clean the email is send to main corporate server and finally send to intended recipient. If the attachment is not clean, ie a steganography payload is detected, alert for data leak can be triggered and that email will not be sent.

In [10] authors present a trustworthiness-based distribution model that aims at data leakage prevention. They study the application where there is a distributor, as a trusted party, managing and distributing files that contain sensitive information to authorized users when they require.In their model, first, the distributor calculates the user's trustworthiness based on his historical behaviors. Then according to the user's trustworthiness and his obtained file set overlapping leaked file set, the distributor accesses the probability of the user's intentional leak behavior as the subjective risk assessment. Then the distributor evaluates the user's platform vulnerability as an objective element. Finally, the distributor makes decisions whether to distribute the file based on the integrated risk assessment. .

3. CONCLUSION

The use of internet for communication purpose has rapidly increased and it magnified the attacks to users. Protecting the data is a big challenge for computer users. The leak of sensitive data on computer systems poses a serious threat to organizational security. Statistics show that the lack of proper encryption on files and communications due to human errors is one of the leading causes of data loss. The threat now extends to our personal lives: a plethora of personal information is available to social networks and smartphone providers and is indirectly transferred to untrustworthy third party and fourth party applications. The data breaches reported are just the tip of the iceberg, as most cases of information leakage go unreported due to fear of loss of customer confidence or regulatory penalties. In this paper a technical survey of recent methods which aims to detect and prevent data leakare presented.

REFERENCES

[1] AsafShabtai, Yuval Elovici and LiorRokach , “A Survey of Data Leakage Detection and Prevention Solutions”, SpringerBriefs in Computer Science .

[2] A Handbook on Information Security Management, Harold F. Tipton, Micki Krause, Fifth Edition.

[3] XiaokuiShu, Danfeng Yao and Elisa Bertino, “Privacy-Preserving Detection of Sensitive Data Exposure”, IEEE Transactions on Information Forensics and Security, 1092-1103.

[4] XiaokuiShu and Jing Zhang, Danfeng Daphne Yao and Wu Chun Feng, “ Fast Detection of Transformed Data Leaks, IEEE Transactions on Information Forensics and Security”, 528-542.

[5] Michael Backes ,Niklas Grimm and Aniket Kate, “Data Lineage In Malicious Enviornments, IEEE Transactions on Dependable and Secure Computing”, 178-191.

[6] P. Papadimitriou, H. Garcia-Molina, “Data Leakage Detection”, IEEE Transactions On Knowledge And Data Engineering, 51-63.

[7] SubhashiniPeneti and B. Padmaja Rani, “Data Leakage Prevention System WithTimeStamp”, International Conference on Information Communication and Embedded Systems, 1-6.

[8] Gilad Katz, Yuval Elovici, and BrachaShapira, “CoBAn: A context based model for data leakage prevention”, Information science on Springer.

[9] VeronikiStamatiKoromina and Christos Ilioudis, “Insider Threats in Corporate Environments: A Case Study for DLP”, in Proc. ACM.

[10] Yin Fan, Wang Yu, Wang Lina, YuRongwei, "A Trustworthiness-Based Distribution Model for Data Leakage Prevention", Wuhan university journal of natural sciences,2010, Vol.15 No.3, 205-209.