American Journal of Engineering Research (AJER) 2016
American Journal of Engineering Research (AJER)
e-ISSN: 2320-0847 p-ISSN : 2320-0936
Volume-5, Issue-2, pp-82-87
www.ajer.org
Research Paper Open Access

Survey on Techniques for Detecting Data Leakage

Bhosale Pranjali A. Gore Anita B. Pandit Sunita K. Prof. Amune Amruta C.
Department of CSE, G.H.Raisoni College of Engg. & Management, Ahmednagar, Savitribai Phule Pune University, India

ABSTRACT: *In current business scenario, critical data is to be shared and transferred by organizations to many stake holders in order to complete particular task. The critical data include intellectual copyright, patient information etc. The activities like sharing and transferring of such critical data includes threats like leakage of information, misuse of data, illegal access to data and/or alteration of data. It is necessary to deal with such problem efficiently and effectively, popular solutions to this problem are use of firewalls, data loss prevention tools and watermarking. But sometimes culprit succeeds in overcoming such security measures hence, if organizations becomes able to find out the guilty client responsible for leakage of particular data then risk of data leakage is reduced. For this many systems are proposed, this paper includes information about techniques discussed in some of such methodologies.*

Keywords: *Client/server, watermark, access right, authentication, encryption, decryption, data leakage, fake records.*

I. INTRODUCTION

A data distributor has given sensitive data to a set of supposedly trusted agents (Third parties) some of the data is leaked and found in an unauthorized place. This problem is known as data leakage problem. To identify the culprit who has leaked the critical organizational data is major challenge. This raises the risk of private data falling into unauthorized hand, whether caused by malicious intent, or an inadvertent mistake, by an insider or outsider, exposed sensitive information can seriously hurt an organization.

Traditionally, watermarking i.e. the process of embedding unique code in each distributed document is used to handle data leakage detection. If the watermarked copy is revealed in the hands of unauthorized party then such party will be declared as culprit. But every time it is not necessary that administration get to know about leakage or it can take time to server to know about the same after analysing such drawbacks of existing systems many researchers proposed different solutions for detection of data leakage.

II. RELATED WORK

2.1 TECHNIQUES FOR DATA LEAKAGE IDENTIFICATION:

2.1.1 Data leakage detection: A survey [1] by Sandip A. Kale C., Prof.S.V. Kulkarni C, 2012

The model proposed in this paper introduces unobtrusive technique for detecting leakage of a set of object or record [1]. This model is proposed for assessing guilt of agent.

The algorithms stated are as follows:

A. Evaluation of explicit data request algorithms:

This algorithm is used for identifying whether fake objects that improves the chances of detecting the culprit.

B. Evaluation of sample data request algorithms:

The distributor is "forced" to allocate certain objects to multiple agents only if the number of requested objects exceeds the number of requested objects in set T [1].

For successful identification of guilty agent 5 modules are declared in this paper, first of them is data allocation module which works for intelligently distributing the data by admin so as to improve probability of detecting the guilty client. Then the fake object module adds fake objects (objects generated by admin altering original document) to data this use of fake objects is inspired by the use of 'trace' records in mailing lists. The optimization module is the distributor's data allocation to agent which has constraints to satisfy agents request and objective is to recognize leakage. The data distributor model enables admin to view which file is leaking. The fifth model i.e. agent guilt module estimates probability of particular data guessed by target and improves the chances of indentifying guilty client.

2.2.2 Detection of guilty agents [2] by S.Umamaheswari, H.Arthi Geetha, 2011

The problem defined in this research work is "The distributor's data allocation to agents has one constraint and one objective. The distributor's constraint is to satisfy agents' requests, by providing them with the number of objects they request or with all available objects that satisfy their conditions. His objective is to be able to detect an agent who leaks any portion of his data.''

The two types of requests are used here:

Sample request Ri = SAMPLE (T; mi): Any subset of mi records from T can be given to Ui and Explicit request Ri = EXPLICIT (T; condi): Agent Ui receives all the T objects that satisfy condition, determined by user provides them data accordingly.

The modules described in this system are:

A. Database maintenance: Here details of agents and data requested by them are maintained.
B. Agent maintenance: In the first part i.e. registration the information about agent is collected and in second part i.e. history the entire details and transactions of users are maintained.
C. Detecting guilty agent: The goal of this module is to estimate the likelihood that the leaked data came from the agent as opposed to other sources, so that guilty agent cannot prove his self innocent.
D. Data allocation: Here, according to two requests handled (sample and explicit) and use of fake objects in data distribution four problem instances are generated and they are EF, EF, SF and SF.

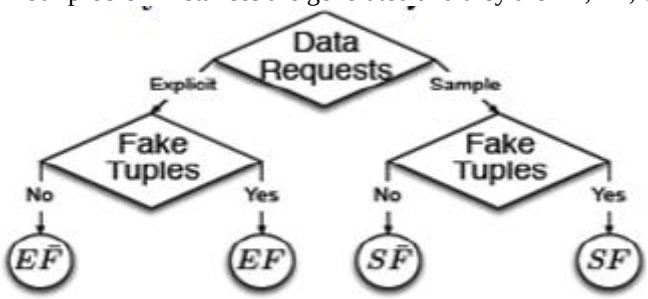

Fig 1. Leakage problem instances [3]

E. Adding fake objects: The addition of fake objects is done on data to be distributed so as to improve the effectiveness in identifying the culprit. The files with some modifications done by administrators are maintained these files are called as trace files which helps to identify improper use of data.
F. Database design: This module maintains all the records in particular manner.

The implementation of this system possible on Netbeans IDE, in input module login, registration etc form provided to user and user has to fill information accordingly then at output part identification and validation of user is done and also guilty agent is detected.

2.2.3 Data leakage and detection of guilty agent [3] by Rupesh Mishra and DK Chitre, 2012

There are lots of other works for avoiding data leakage like watermarking and mechanism that allow only authorized user's access sensitive data through various access control policies.

The system in this work has strict constraints where the distributor may deny to serve an agent and may not provide agents with perturbed versions of same objects and the objective is provided so as to maximize the chances of recognizing guilty agent. The data allocation strategies has given prime importance, because if the data is distributed properly then culprit can easily traced out also implicit and explicit request problem are handled in data allocation strategy.

A distributor owns a set T = {t1, t2, ..., tm) of valuable object and agent U1, U2, ... , Un. An agent Ui receives a subset of objects Ri subset T, either by sample request or by explicit request. The event that user Ui is culprit and leaked data s denoted by {Gi | S}, whereas Pr{Gi | S} is probability that agent Ui is culprit given evidence S. For the distribution the algorithm estimated is as follows:

Algorithm 1: Distribution Algorithm

1. **Assignment:** Assigning DID to data objects.
2. **Hashing:** H(AID)→ DID
3. **Fake record generation:** FID and Fake Record
4. **Mapping:** FID→ DID with P.K. of record
5. **Backup & Removal:** Store information and remove DID.

And once the object is distributed the guilty agent detection is performed by following algorithm:

Algorithm 2: Agent Detection

Set **S T** is obtained at unauthorized place

1. **Mapping DID**

2. **Agent Tracing**
    1. Distribution ID set
    2. Fake record
    3. Missing record
3. **Estimate guilt.**

2.2.4 Data leakage detection [4] by Sandip A. Kale, Prof. S.V. Kulkarni, 2012

This paper introduces 'unobtrusive' technique for detecting data leakage of set of object. The various techniques like watermarking are studied like embedding and extraction, secure spread spectrum, DCT based watermarking, spread spectrum, wavelet based watermarking, robust watermarking technique, invisible watermarking etc.

Data allocation module allows admin to send the files to authenticated client, the fake object module adds altered objects to database and they are distributed to increase the probability of detecting culprit. Optimization module has constraint to satisfy distributors request and objective to identify guilty client. According to experimental results of agent guilt module the probability that agent Ui is guilty is:

$$\Pr\{Gi \mid S\} = 1 - \prod_{te S \cap Ri} (1 - \frac{1-p}{|Vt|})$$

Eqn 3[4]

2.2.5 Detection of data leakage in cloud computing environment [5] by Neeraj Kumar, Vijay Katta, Himanshu Mishra, Hitendra Garg, 2014

The technique proposed focuses on how data leakage and data leakager is identified in cloud computing environment, also concentrates on how more proposed work uses the Bell-LaPadula model to analysis and design of secure computer system. The notion of "secure state" is defined and it is proven that each state transition preserve security by moving to other secure state, thereby inductively proving that system satisfies security objective of the model [5]. In the Bell-LaPadula model each subject S has a lattice of rights and access rights provided by it are reading down (NRU), writing up (NWD), simple security property, star property, read only, append only, execute only, read-write etc.

The first stage of the model described here is registration of client to server, after this server maintains database about the client which is called as server directory table and some fields of this table are client Id, SHA-512(hash) and (m, n). The next step is watermarking of secret message this technique is implemented using 2 steps in first the place where the watermarking should be done is calculated:

- Row positioning pixel, m = I (1; 1) + 2
- Column positioning pixel, n = I (1; 2) + 2

The secret message, encryption key K, AES-128 are used to calculate cipher text C, for placement of cipher C and authentication code M pixel value of image starting from (m, n) in original document is replaced by cipher C. Now the watermarked document is sent to client this process is shown in figure given below:

Fig 2. Sending watermarked document to client

To detect client id reverse process is applied i.e. first the placement of cipher C and authentication code M in document is pointed out for this server uses table where the point (m, n) is stored. Once the authentication code M is found the secret message IDc is calculated and verified for that AES- 128 decryption method is used.

2.2.6 Detecting Data Leakage in Cloud Computing Environment (A Case Study of General Hospital Software) [6] by Alex Ofori Karikari, Joseph Kobina Panford, James Ben Hayfron-Acquah, Frimpong Twum

For the development of system first analysis of old hospital system was conducted and the searching was that, the system does not have any transaction log so anything entered or deleted from the system could not be traced by administrator, and also analysing the existing system problem statement defined as the data leakages are a big concern to organisation as well as individuals and also costing huge sums of money to institutions.

The purpose of proposed methodology was to show how user accessibility and activity to a computing resource could be traced, monitored and audited to safeguard the integrity, accessibility and availability of data to authorized and authenticated users in cloud application. The proposed system has transaction log to monitor user activities, biometric verification system is used for client authentication purpose. Simulation test was done with the dual purpose how a wrongly configured network topology could grant access to both known and unknown user (data clerk) into the system to breach/leak data for instance [6]. This system allows administrators track down at least users in their system.

2.2.7 Data leakage analysis on cloud computing [7] by Bijayalaxmi Purohit, Pawan Prakash Singh, 2013

In this paper cloud describes the use of collection of services, applications, information, and infrastructure. Services like computation, network, and information storage. The major areas of focus are: - Information Protection, Virtual Desktop Security, Network Security, and Virtual Security. The need to protect such a key component of the organization cannot be over emphasized. Data Loss Prevention has been found to be one of the effective ways of preventing Data Loss. DLP solutions detect and prevent unauthorized attempts to copy or send sensitive data, either intentionally or unintentionally, without authorization, by people who have authorized access the sensitive information. Data loss, which means a loss of data that occur on any device that stores data. In this paper, we deal with both the terms data loss and data leakage in analyzing how the DLP technology helps minimizing the data loss/leakage problem? DLP technology minimizes the data loss problem in the organization. Data Loss/Leakage Prevention (DLP) is computer security which is used to finding, monitor, or protect data in use, data in motion, or data at rest. DLP is used to identify sensitive content by using deep information analysis to per inside files or with the use if network communications. DLP is mainly designed to protect information assets in minimal interference in the business processes. It also enforces protective controls to prevent unwanted incidents. DLP can be used to reduce risk and to improve data management practices and even lower compliance cost. Systems are designed to detect and prevent unauthorized use and transmission of confidential information.

2.2.7.1 Classification of Information Leakage:

In this paper information leakage into three levels which means a document containing confidential data can be classified as the unintentional leak, intentional leak, and malicious leak. Activities under unintentional Leak are:

1. Attach document
2. Zip and send
3. Copy & Paste

The unintentional leakage normally occurs when a user mistakenly sends a confidential data and information to third party or wrong recipient. This is done without any personal intention. For instance, if an employee sends an email attaching document mistakenly this contains confidential data to a wrong person.

Intentional Leak:

The intentional leakage normally occurs when user tries to send a confidential document without aware of company policy or finally sends anyhow. This is usually done when a user bypassing the security rules and regulations or devices without trying to gain personal benefits. For instance, when an employee renames a document folder and partially copies data from it. Intentional Leak can be performed using following activities:

1. Document renames
2. Document type change
3. Partial data copy
4. Remove keyword

Malicious Leak:

Malicious leakage usually caused when user deliberately trying to sneaks the confidential data past security rules. Malicious Leakage is done using:

1. Character encoding
2. Print screen
3. Password protected
4. Self extracted archive
5. Hide data
6. Policies or product.

For instance, when an employee sneaks a confidential data from enterprise system and sends them through email and even cause vulnerability to the system. In this paper, we do analysis of data leakage prevention. Why it can balance the data security and user convenience.

2.2.8 Detection of Data Leakage Using Unobtrusive Techniques [8] by Mr. Ajinkya S. Yadav, Mr. Ravindra P. Bachate, Prof. Shadab A. Pattekari, 2013.

Demanding market conditions encourage many companies to the outsource certain business processes (e.g. marketing, human resources) and associated activities to third party. This model is referred as Business Process Outsourcing (BPO) Security and business assurance are essential for the BPO.

In most cases, the service providers need access to a company's intellectual property or other confidential information to carry out their services. For example human resources BPO vendor may need access to the employee databases with sensitive information. The main security problem in BPO is that the service provider may not be fully trusted or may not be securely administered. Business agreements for BPO try to regulate how the data will be handled by service providers, but it is almost impossible to truly enforce or verify such policies across different administrative domains.

2.2.8.1 Unobtrusive Techniques:

In this develop a model for assessing the "guilt" of agents, also present algorithms for distributing objects to the agents, in a way that improves our chances of identifying a leaker. The distributor may be able to add fake objects to distributed data in order to improve his effectiveness in detecting guilty agents. However, fake objects may impact correctness of what agents do, so they may not always be allowable. Consider the option of adding "fake" objects to distributed set. Such objects do not correspond to real entities but appear realistic to agents. In a sense, the fake objects acts as a type of watermark for the entire set, without modifying the any individual members. If it turns out an agent was given one or more fake objects that were leaked, then distributor can be more confident that agent was guilty.

III. PERFORMANCE ANALYSIS

3.1 Accuracy for finding guilty agent:

The systems mentioned in this paper except "Detection of data leakage in cloud computing environment" does not provide much accuracy for finding guilty agent, hence we are preferring the technique of embedding the secret key in the data requested by client, as this key is unique for every data and client the accuracy and chances of indentifying culprit can be increased. For the encryption of key we prefer AES-128 algorithms.

The reason why AES-128 preferred over DES and RSA algorithm is given by following tables which was drawn after analysing [7],[8].

Sno	DES	AES	RSA	Data Size
1	3.0	1.6	7.3	153KB
2	3.2	1.7	10.0	118KB
3	2.0	1.7	8.5	196KB
4	4.0	2.0	8.2	868KB
5	3.0	1.8	7.8	312KB

Table1. Comparison of various packet sizes for DES, AES & RSA algorithm (Encryption Time) [5]

Sno	DES	AES	RSA	Data Size
1	1.0	1.1	4.9	153KB
2	1.2	1.2	5.0	118KB
3	1.4	1.24	5.9	196KB
4	1.8	1.2	5.1	868KB
5	1.6	1.3	5.1	312KB

Table 2: Comparison of various packet sizes for DES, AES & RSA algorithm (Decryption Time) [5]

Factor Analyzed	DES	AES	RSA
Development Years	1977	2000	1978
Key-Length (Bits)	56	128,192,256	$\leq$1024
Nature of Algorithms	Symmetric	Symmetric	Asymmetric
Encryption/Decryption(Speed)	Low	High	Medium
Nature of Security Attacks	Inadequate	Highly Secured	Highly Secured

Table 3: Analysis of various factors [5]

3.2 Watermarking techniques:

There are various watermarking systems are available like secure spread spectrum watermarking, wavelet based watermarking, robust watermarking, DCT-based watermarking etc. But we prefer "Invisible watermarking" for embedding the key because of its following features:

A. As watermarking is performed in most significant region of data, if culprit tries to remove or destroy the watermark it will degrade the appearance quality of data which helps to identify misuse of data.
B. The creation of watermark using input user watermark (logo) is allowed.
C. Preserves quality of host image.
D. Allows robust-insertion and extraction of watermark.
E. Ownership proof could be established under hostile attacks [4].

IV. CONCLUSION:

In this paper, we have studied different papers about data leakage that takes to conclusion that though there techniques like watermarking were available for securing data but these were inefficient, also the system stated are consists of fake records, explicit and implicit requests etc these can cause overhead to system so we are proposing a system which requires less memory and it will be affordable to low budget organisations also.

REFERENCES

[1] Sandip A. Kale C1, Prof.S.V. Kulkarni C "Data Leakage Detection: A Survey" IOSR Journal of Computer Engineering (IOSRJCE) Vol.1, Issue 6 (July-Aug 2012), PP 32-35.
[2] S.Umamaheswari, H.Arthi Geetha "Detection of Guilty Agents, Proceedings of National Conference on Innovations in Emerging Technology-2011.
[3] Rupesh Mishra and DK Chitre. "Data leakage and detection of guilty agent". International Journal of Scientific & Engineering Research, 3(6), 2012.
[4] Sandip A. Kale, Prof. S.V. Kulkarni" Data leakage detection" International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 9, November 2012.
[5] Neeraj Kumar, Vijay Katta, Himanshu Mishra & Hitendra Garg, "Detection of Data Leakage in Cloud Computing Environment", in Sixth International Conference on Computational Intelligence and Communication Network, 2014.
[6] Alex Ofori Karikari, Joseph Kobina Panford, James Ben Hayfron-Acquah, Frimpong Twum "Detecting Data Leakage in Cloud Computing Environment (A Case Study of General Hospital Software)" International Journal of Scientific Engineering and Applied Science (IJSEAS) - Volume-1, Issue-3, June2015 ISSN: 2395-3470.
[7] Bijayalaxmi Purohit, Pawan Prakash Singh "Data leakage analysis on cloud computing" International Journal of Engineering Research and Applications (IJERA) Vol. 3, Issue 3, May-Jun 2013, pp.1311-1316.
[8] Mr. Ajinkya S. Yadav, Mr. Ravindra P. Bachate, Prof. Shadab A. Pattekari "Detection of Data Leakage Using Unobtrusive Techniques", IOSR Journal of Computer Engineering (IOSR-JCE) vol. 8, Issue 4 (Jan. - Feb. 2013), PP 79-84.
[9] NIST FIPS Pub. 197. Announcing the Advanced Encryption Standard (AES), 2001.
[10] Jakob Jonsson and Burt Kaliski. Public-key cryptography standards (pkcs)# 1: Rsa cryptography specifications version 2.1. 2003.