THEFT IDENTIFICATION DURING DATA TRANSFER IN IT SECTOR

C.preethi
Assistant Professor,Department of CSE
Veerammal Engineering College
Dindigul,Tamilnadu
Email.ID: preethic20@gmail.com

R.Ramesh
Assistant Professor,Department of CSE
Veerammal Engineering College
Dindigul,Tamilnadu
rameshrcse@gmail.com

***Abstract*- Theft identification during data transfer can be elaborated as in when a data distributor has given sensitive data to the trusted agents and some of the data is leaked and found in an unauthorized place. For this the system can use data allocation strategies or can also inject "realistic but fake" data records to improve identification of leaked data and who leaks the data. The Fake Objects looks exactly like original data in which the agents cannot be identified. Many of the data from the organization can be mostly leaked through the e-mails. So we have to filter those E-mails from the organization. In order to secure the data which are leaked from the mail can be detected and identified through the "Fake Objects". E-Random and S-Random algorithm are used to minimize the content as well as to detect the guilty agent. The leaked data from the organization can be sent to the third parties in the form of cipher text.**

Keywords: Data leakage, Fake object, E-random, S-random, E-mail filtering system.

I.INTRODUCTION

We consider applications where the original sensitive data cannot be perturbed. Perturbation is a very useful technique where the data are modified and made "less sensitive" before being handed to agents. For example, one can add random noise to certain attributes, or one can replace exact values by ranges. However, in some cases, it is important not to alter the original distributor's data. For example, if an outsourcer is doing our payroll, he must have the exact salary and customer bank account numbers. If medical researchers will be treating patients (as opposed to simply computing statistics), they may need accurate data for the patients. Type of data being leaked through e-mail can be in the form of graphical files, video files, text documents, audio files, zip files, etc. In the proposed system, the owner of the data is the distributor and the supposedly trusted parties are the agents. The main of the system is to detect the leak of distributor's sensitive data that must have been leaked by agents. And also to convert the data from the plain text to the cipher text which are leaked through the mails of the guilty agent.

II. BASIC SURVEY

In the existing system, watermarking technique is used to identify where the data leaked from the organization. Watermarking is the technique has the unique code. These unique codes is inserted into each data or records which is then distributed to the clients by the user. If any particular client leaks the given data to the third parties i.e. unauthorized users, then this leaked data and the leaker can be identified by the means of this watermarking technique. Watermarking is very useful technique but it has some disadvantages because of which there is a great chances of getting the data leaked. Digital watermarking technique is also there in which the code is embedded in the digital file like audio or video files. If the watermarking technique is used then the code which is embedded in the data can be modified because of which it becomes very difficult to identify the guilt agent or leaker. Furthermore these watermarks can be destroyed if the data recipients are malicious. E.g. data of private firms, corporate sectors or the research centers. A research center may require a patient's data from the hospitals to do some research work. Similarly a private firm may have its highly confidential data of its financial statements, employees records etc. Another enterprise may outsource its data processing, so data must be given to various other companies. We call the owner

of the data the distributor and the supposedly trusted third parties the agents. In this paper, we study unobtrusive techniques for detecting leakage of a set of objects or records. Specifically, we study the following scenario: After giving a set of objects to agents, the distributor discovers some of those same objects in an unauthorized place. (For example, the data may be found on a website, or may be obtained through a legal discovery process.) At this point, the distributor can assess the likelihood that the leaked data came from one or more agents, as opposed to having been independently gathered by other means. Using an analogy with cookies stolen from a cookie jar, if we catch Freddie with a single cookie, he can argue that a friend gave him the cookie. But if we catch Freddie with five cookies, it will be much harder for him to argue that his hands were not in the cookie jar. If the distributor sees "enough evidence" that an agent leaked data, he may stop doing business with him, or may initiate legal proceedings.

III. PROPOSED SYSTEM

Due to the disadvantages of the existing system it is not possible to detect the guilty agent. So it is difficult to find who leaks the data. The main aim of the proposed system is to find who leaks the data and where the data leaks. In the proposed system, it is going to implement the concept of "Fake Objects". Now if suppose the director of the company wants to share some sensitive data (records) with clients of his company but he does not want his data to be leaked anywhere in between. So before sending the sensitive data to the clients what the proposed system will do is it will add fake objects (record) in database which will exactly look like original data. The Client will be unaware of these fake objects. Only the director of company knows that where and how many fake objects are inserted.

As well as we create the fake objects to each and every persons mail-id in the organization. When the guilty agent leaks the data, it is easy to find out the agent using those fake objects. The message sent to the third parties is in the cipher text form. So the original data will not be leaked.

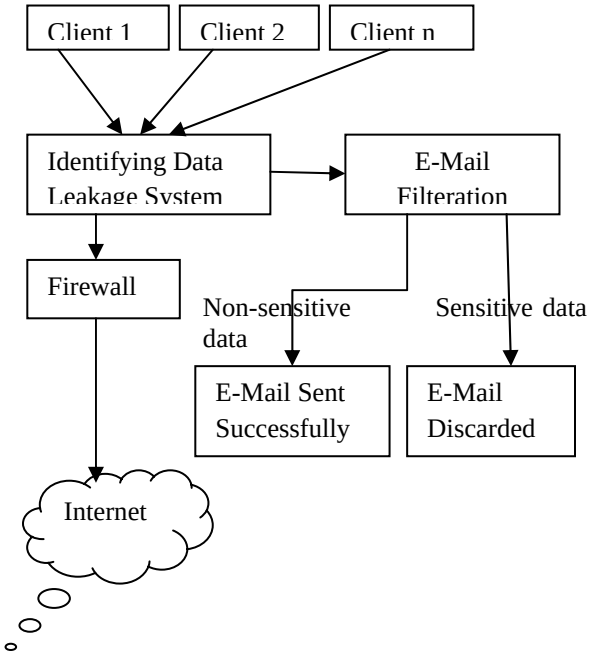

Fig 1: Divulging Data Leakage Detection

A. DATA ALLOCATION PROBLEM

The main focus of this paper is the data allocation problem: how can the distributor "intelligently" give data to agents in order to improve the chances of detecting a guilty agent? As illustrated in Fig. 2, there are four instances of this problem we address, depending on the type of data requests made by agents and whether "fake objects" are allowed. The two types of requests we handle were defined in sample and explicit. In sample data request agent receives a subset of distributers data which required by agent. In explicit data request data satisfying a specific condition is given to agent.

B. FAKE OBJECTS

The distributor may be able to add fake objects to the distributed data in order to improve his effectiveness in detecting guilty agents. However, fake objects may impact the correctness of what agents do, so they may not always be allowable. The idea of perturbing data to detect leakage is not new. However, in most cases, individual objects are perturbed, e.g., by adding random noise to sensitive salaries, or adding watermark to an image. In our

case, we are perturbing the set of distributor objects by adding fake elements. In some applications, fake objects may cause fewer problems that perturbing real objects. For example, say that the distributed data objects are medical records and the agents are hospitals. In this case, even small modifications to the records of actual patients may be undesirable. However, the addition of some fake medical records may be acceptable, since no patient matches these records, and hence, no one will ever be treated based on fake records. Our use of fake objects is inspired by the use of "trace" records in mailing lists. In this case, company A sells to company B a mailing list to be used once (e.g., to send advertisements). Company A adds trace records that contain addresses owned by company A. Thus, each time company B uses the purchased mailing list, A receives copies of the mailing. These records are a type of fake objects that help identify improper use of data.

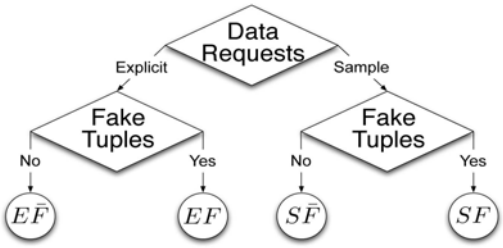

Fig 2: Data Leakage Problem

C. E-RANDOM ALGORITHM

E-random algorithm is also called as Explicit data request algorithm. In case of explicit data request with fake not allowed, the distributor is not allowed to add fake objects to the distributed data. So Data allocation is fully defined by the agent's data request. In this algorithm, the agent receives the entire data object that satisfies the condition of the agents' data request. In case of explicit data request with fake allowed, the distributor cannot remove or alter the requests R from the agent. However distributor can add the fake object. In algorithm for data allocation for explicit request, the input to this is a set of request from n agents and different conditions for requests. The e-optimal algorithm finds the agents that are eligible to receiving fake objects. Then create one fake object in iteration and allocate it to the agent selected. The e-optimal algorithm minimizes every term of the objective summation by adding maximum number of fake objects to every set yielding optimal solution.

Algorithm:

E-optimal solution
O(n+n2B)= O(n2B)
Where n= number of agents,
B= number of Fake objects.

D. S-RANDOM ALGORITHM

With sample data requests agents are not interested in particular objects. Hence, object sharing is not explicitly defined by their requests. The distributor is "forced" to allocate certain objects to multiple agents only if the number of requested objects exceeds the number of objects in set T. The more data objects the agents request in total, the more recipients on average an object has; and the more objects are shared among different agents, the more difficult it is to detect a guilty agent. In this algorithm, the agent receives only the subset of data object that can be given to the agent. The working of Sample Data Request algorithm is same as the working of Explicit Data Request.

IV.MODULES

- User Authentication
- Fake Object Generation
- E-Random Implementation
- S-Random Implementation
- Data Distributor
- E-Mail Filtering
- Generation of Cipher Text

A. USER AUTHENTICATION

In this module, user registration process is done by the administrator. Here every user will give their details for registration. User details are encrypted by using AES (Advanced Encryption Standard) Algorithm. Authenticated person only can register in this process. Administrator will generate a username and password for every user. The user can use that username and password for login process.

B. FAKE OBJECT GENERATION

In this module, the owner of the data or information will add some fake objects (record) in database which will exactly look like original data. The client will be unaware of these fake objects. Only the owner of the data knows where and how many fake objects inserted into original data.

C. E-RANDOM IMPLEMENTATION

The agent receives the entire data object that satisfies the condition of the agents' data request. In case of explicit data request with fake allowed, the distributor cannot remove or alter the requests R from the agent. However distributor can add the fake object. The e-optimal algorithm minimizes every term of the objective summation by adding maximum number of fake objects to every set yielding optimal solution.

D. E-OPTIMAL SOLUTION

$$O(n+n2B) = O(n2B)$$

Where,

n= number of agents,

B= number of Fake objects.

E. S-RANDOM IMPLEMENTATION

In this module the more data objects the agents request in total, the more recipients on average an object has; and the more objects are shared among different agents, the more difficult it is to detect a guilty agent. In this algorithm, the agent receives only the subset of data object that can be given to the agent. The working of Sample Data Request algorithm is same as the working of Explicit Data Request.

F. DATA DISTRIBUTOR

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 (e.g., on the web or somebody's laptop). The distributor must assess the likelihood that the leaked data came from one or more agents, as opposed to having been independently gathered by other means.

G. E-MAIL FILTERING

In this module, there are two possible ways by which an agent can leak the data. First one is using storage devices like pen drive or hard drives and second is by using email. First way will be handled by Data leakage detection module and for second way are going to use e-mail filtering module. In e-mail filtering module when client will attach database which he wants to leak through his e-mail address, before sending data will perform check on that attachment. And if we sensitive data then the system will discard that mail. Agent will not be able to send sensitive data through e-mail. Algorithm used for e-mail filtering is K nearest neighbor.

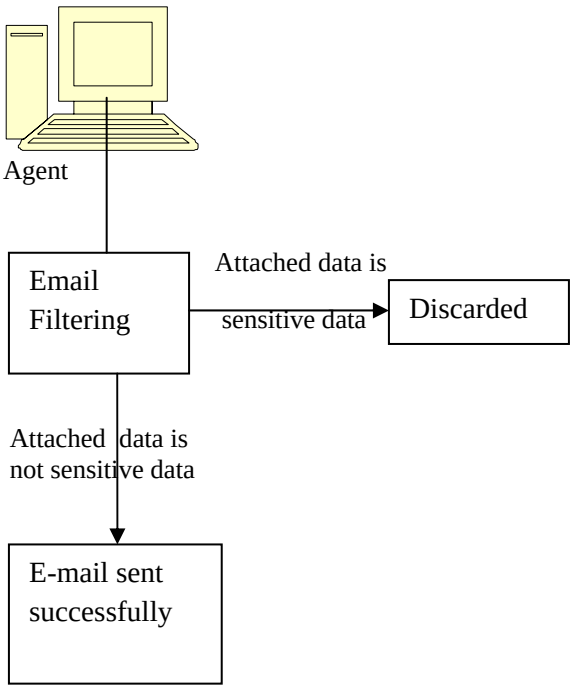

Fig 3: Email filtering system

It flow is as follows:

1. Identify the data.
2. Remove stopping words such as this, is, a, etc.
3. Remove or change the synonyms.

4. Calculate the priority of the word depending upon the sensitivity of the data.
5. Compare data with predefine company data sets.
6. Filter the data if it has company's important data sets.

H. GENERATION OF CIPHER TEXT

After sending the mail from the guilty agent to the third party, the mail-id is first verified by the admin. If the mail contain any sensitive data or any others mail-id then the fake objects are used to intimate that the agent is guilty. After finding that the mail is sensitive then the data will be converted from plain text to cipher text. So that the mail received to the third party cannot able to read the data. AES (Advanced Encryption Standard) algorithm is used to encrypt the data in that particular mail-id.

V.CONCLUSION

In the real time system, the data must be very confidential during data transfer and it should not be handed over to the third parties who will unknowingly or maliciously leak it. Even if we had to hand over sensitive data, to the agents, we could use watermarking technique for each object so that we could trace its origins with absolute certainty. But, in many cases, the agents may not be fully trusted, and we may not conclude weather the leaked data or object came from an agent or from some other source, since certain data cannot admit watermarks. If the agent is malicious he can easily remove the watermarks and he can obtain the original data. Because of these difficulties, we have shown that it is possible to assess the likelihood that an agent is responsible for a leak, based on the overlap of his data with the leaked data and the data of other agents, and based on the probability that objects can be "guessed" by other means. Our model is relatively simple, that adds the fake objects into the content as well as we create the fake object for each and every person’s mail-id also. The algorithms we have presented implement a variety of data distribution strategies that can improve the distributor's chances of identifying a leaker. Even if the data is leaked by the agent to the third parties it will be received in the cipher text format. We have shown that distributing objects judiciously can make a significant difference in identifying guilty agents, especially in cases where there is large overlap in the data that agents must receive.

VI.REFERENCES

[1].R. Agrawal and J. Kiernan, "Watermarking Relational Databases, "Proc. 28th Int'I Conf. Very Large Data Bases (VLDB '02), VLDB. Endowment pp. 155-166, 2002.

[2].L. Sweeney, "Achieving K-Anonymity Privacy Protection Using Generalization and Suppression," http: //en.scientificcommons. Org/43 196131,2002

[3]. F. Guo, J. Wang, Z. Zhang, X. Ye, and D. Li, "An Improved Algorithm to Watermark Numeric Relational Data," Information Security Applications, pp. 138-149, Springer, 2006.

[4]. S. Jajodia, P. Samarati, M.L. Sapino, and V.S. Subrahmanian, "Flexible Support for Multiple Access Control Policies," ACM Trans. Database Systems, vo1.l34

[5] B. Mungamuru and H. Garcia-Molina, "Privacy, Preservation and Performance: The 3 P's of Distributed Data Management," technical report, Stanford Univ.,2008

[6].Panagiotis Papadimitriou and Hector Garcia-Molina, "Data Leakage Detection, "lEEE Transactions on Knowledge and Data Engineering, Vol 23, No.1 january2011.

[7] Ankit Agarwal, Mayuir Gaikwad, ”Robust data leakage and E-mail Filtering System” International Conference on Computing, Electronics and Electrical Engineering,[ICCEET],2012.