A Model for Traitor Detection with Appraising Approach

Sai Prasad Kashi
Student, CSE Dept.,
IARE, Dundigal, Hyd.

N. Chandra Sekhar Reddy,
Ph. D
Professor in CSE Dept., IARE,
Dundigal, Hyd.

G. Prabhakar Reddy
Student, CSE Dept.,
IARE, Dundigal, Hyd.

ABSTRACT
In today's technically empowered world it is a major task for distributors to prevent their data from false agents. Data distribution across trusted third party agents is complicated and always in the danger of misconfiscation by the users or agents. Loss of large volumes of shielded information has become regular headline event. Due to this reason data accessing in a secure way is became a hot topic of research and it became a challenging part to identifying leakages. In case of a leak, nothing can be done by the data distributor. Only possibility is forcing companies to re-issue cards, notify customers and mitigate loss of goodwill from negative publicity. The existing methods sometimes fail to do the same. In this work, we develop an algorithm for distributing data to agents, in such a way that improves the chances of identifying a leakage. We consider adding "fake data" objects to the distributed original data which do not correspond to real entities but appear realistic to the agents. Here this type of fake objects worn as a type of watermark for the entire set, without modifying any individual data. If an agent was given one or more fake objects that were leaked, then the distributor can find the corrupted agent who leaked the data. We consider adding "fake data" objects to the distributed original data which are not equivalent to real entities but appear as a practical one to the agents. If an agent was given one or more fake objects that were leaked, then the distributor can identify the guilty agent.

General Terms
Data leakage, Security, fake identification, authentication et. al.

Keywords
Data leakage, guilty agent, Traitor, Fake data.

1. INTRODUCTION
Data leakage detection is measured as an important feature in business dealings primarily in present drift. Organizations share customer valuable data with other companies who are in association with that company. In this circumstances data protection is important so data leakage discovery will play significant role. Data leak is the "unofficial or involuntary exposure, revelation, or loss of sensitive information". Many businesses have in their control delicate data about their organization, employees and customers [1].

Main goal of this paper is to detect the traitor who leaked the data. Conservatively, water marking is used for detecting leakages. If the recipient is vicious he may destroy the watermark and access control over the shielded data. It also place limitations on clients so our aim is to provide service to all clients. Later techniques as Perturbation theory is used to modify the data and makes it less delicate before being handed to agents and use of unobtrusive techniques for traitor detection [3].

Let us consider a small issue:

Whenever an agent has given set of objects and in many cases as of today if the distributor finds some of those same objects in unlawful place. The data leak may be through any illegal discovery process. Then the distributor should be confident to identify one's own data and assess likelihood that the leaked data came from one or more agents. In detail let us consider there are 'n' objects in a set and unknowingly if some objects are missing from set and when the traitor is caught with single evidence it may not provide exactness as the traitor can escape saying it was given by his friend [3]. Hence to catch the traitor and proceed legally multiple evidences are required. Our new approach improves the chances of finding the leakages and traitor compared to that of data.

The data leakage prevention based on the trustworthiness is used to assess the trustiness of the customer. Maintaining the log of all user' request to the data provenance problem. In both the commercial and defense sectors a compelling need is merging for rapid, yet secure, circulation of information. In this paper we address the threat of information leakage that often accompanies such information flows. We focus on domains with one information source (sender) and much information sinks (recipients) where:

i. leaking a shared information is beneficial to the recipients but undesirable to the sender
ii. sharing is mutually beneficial for the sender and the recipients, and
iii. Information sharing decisions of the sender are determined using imperfect monitoring of the unintended information leakage by the recipients [1].

Finally, there are also different mechanisms to allow only authorized users to access the sensitive information through access control policies, but these are restrictive and may make it impossible to satisfy agent's requests.

2. LITERATURE SURVEY
2.1 Motivations
In the past few years, there has been a sharp increase in data leakage by traitors from many organizations. The survey of 2005-2011 by FBI indicates that, Data leakage is the greatest source of financial loss of many organizations. As internet has become a big boom in present years data

leakage has became very easy so above issues motivated me to choose this topic and work for it.

2.1.1 Leak in Media's perspective:

The media is the most fascinating ways of communicating issues worldwide. Data leakage appears to be increasingly more popular in the media as the reported breaches enlarge [2].As per ICO survey there were more than 500 institutions which tested about safety among valuable data as per 2010. This validation supports the theory of there being an increase in breaches during the decline but what must be taken into consideration is that there is an increase in the accounted cases. It may be that more trades are becoming aware of data escape where in earlier times they were insensible to breaches devoted. Reported in the media, a nationwide employee's workstation was stolen from their address containing confidential data. Almost eleven million customers nationally were at risk of uniqueness crime at the time. The Financial Services Authority was watchful by the breach and it was found that the all over the country did not start an analysis until three weeks after the thievery took place. The firm was fined a huge compensation by the city supervisory body for the security desecration [2].

2.2 Objectives

Main aim of this paper is:

1. To detect the leakage of data while transmitting the data from distributor to agent using fake objects.
2. Identifying the guilty agents or traitors in multiple ways.
3. And hence avoiding the leak of valuable data.

2.3 Existing System

Watermark Feature:

A well known tool for including various types of text data in digital content is Digital Watermarking. In common, information for protecting copyrights and proving the validity of data is entrenched as a watermark [3]. The watermark content may be a text document, a video clip, an image or some form of digital data that the owner may look into. The main purpose of the watermark is to identify who the owner of the digital data is, but it can also identify the anticipated recipient. Why do we need to embed such information in digital content using digital watermark technology? The reasons behind emerging of this technology are the boom in the Internet and its wide ranged usage all over the world. It has become easy to connect to the Internet from home computers and provide information using the World Wide Web. The digital date may be encrypted using RSA or using any public keys as shown in fig: 1

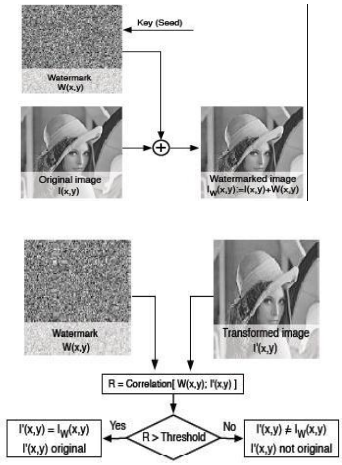


Fig: 1 Watermark encryption and decryption

Complete data handled on the web is provided as digital content. From fig:1 we can see that digital content can be easily copied in a way that makes the new file impossible to differentiate from the original. Then the content can be duplicated in large quantities. To avoid this, currency and stock permit contain watermarks as such as stamps [3]. These watermarks are one of the methods for preventing counterfeit and unlawful use. Digital watermarks apply a similar method to digital content. Watermarked content can prove its origin, thereby protecting copyright. A watermark also discourages piracy by silently and psychologically deters criminals from making unlawful copies.

Problems with Existing System:

Current watermarking approach can detect the traitor but the total number of evidence will be less and they may not be able to proceed legally for further proceedings due to lack of good amount of evidence and the chances to escape of traitors are high, moreover a malicious traitor can break it easily.

If the Traitor is expert and vicious then he can easily break the water mark using many techniques. Chances of identifying the exact culprit is very low, hence advanced algorithms/techniques are required so that the traitor can be caught with many evidences.

2.4 Proposed System

After understanding problems from above issues, in this system the leakage of data is detected by generating fake objects. Our model is relatively simple, but we believe it captures the essential trade-offs. The appraising approach algorithm that we have linked up use a variety of data distribution strategies which improves the distributor's chances of identifying a traitor. Here we are implementing the watermarks which are already implemented along with fake data objects so that the Traitor may not have a proper idea of how to escape, so more chances for catching the guilty agent The below fig:2 depicts the classification of data sets.

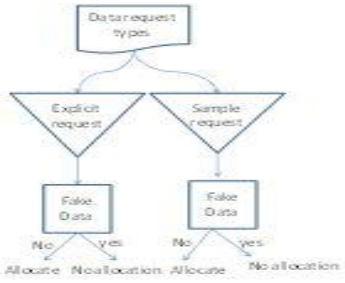


Fig: 2 Classification of data sets

3. SYSTEM DESIGN

This architecture has been implemented differently for detecting the faults but the steps involved here are faraway different by making some changes are made viz., random algorithm section is included in Fig: 3 [9].

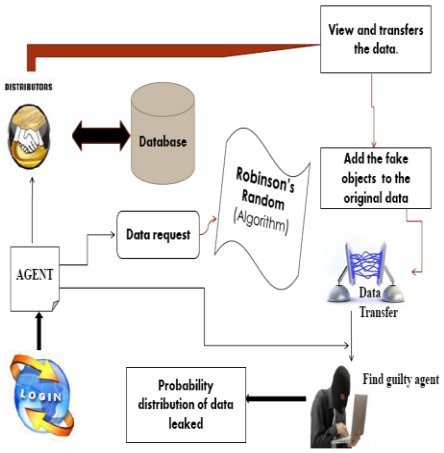


Fig: 3 Architectural view of traitor detection

4. PERFORMANCE ANALYSIS

The project explained in this paper is completely based on the idea of identifying the traitor who is the main source of the data leak. Here, the main objective of this project is to add multiple fake objects to the formerly distributed data in order to find out the traitor. It comprises 4 modules as discussed below:

Data Allocation

The main focus of this paper is the data allocation problem as how can the distributor intelligently give data to agents in order to improve the chances of detecting a guilty agent. In this module, administrator has to login with his id and password. Administrator has all the agent information, user data inside his database.

Administrator is now able to view the database consisting of the original data as well as the fake data. Administrator can also list the agents here. He will be able to add additional information to the database. Agent's information can be added here [5].

Fake Object module

Fake objects are objects generated by the distributor in order to increase the chances of detecting agents that leak data. The distributor may be able to add fake objects to the distributed data in order to improve his effectiveness in detecting guilty agents [5]. Our use of fake objects is inspired by the use of trace records in mailing lists.

Optimization

The Optimization Module is the distributor's data allotment 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.

Data Distribution

A data dispenser has given susceptible data to a set of evidently trusted agents. Some of the data is leaked and found in an unofficial place. The distributor must assess the likelihood that the leaked data came from one or more agents, as opposed to having been independently grouped by other means [5][6].

4.1 Allocation Methodologies

The distributor distributes the data to agents such that he can easily detect the culprit in case of any leak. To improve the chances of detecting traitor, distributor inserts fake objects into the distributed dataset K. These multiple fake items are created in such a manner that, agent cannot distinguish it from original data [7]. In this paper we have used the dataset of fake tuples. Depending upon the addition of fake tuples according to the agent's request, data allocation problem is divided into four types as:

i. Implicit request with fake tuples.

ii. Implicit request without fake tuples.

iii. Explicit request with fake tuples.

iv. Explicit request without fake tuples

So to minimize the overlap, we insert the fake objects in to one of the agent's dataset. In this paper, we presented the **algorithm** and the corresponding results for the explicit data allocation with the addition of fake tuples. Whenever any user request for the tuples, it follows the following steps:

Algorithm:

Step I: The **request(R)** is sent by the user to the **distributor (D).**
Step II: The request may be **implicit (Ri) or explicit (Re).**
Step III: If **R = Ri** then a subset of the data is given.
Step IV: If **R= Re**, it is checked with the log, if any previous request is same.
Step V: If **R=same** then system gives the data objects that are not given to previous **agent (A).**
Step VI: The **fake objects (Fi)** are added to agent's request set.
Step VII: **Leaked data set (L),** obtained by distributor is given as an input.
Step VIII: Calculate the **guilt probability (G)** of user.

Algorithm for reading a file:

I. Input the file which you want to read F
II. Create object for HWPF type
III. PgNumber==1
IV. Create an object for word extractor "WE"
V. Use "WE" object for getting the file F.

Algorithm for reading the contents of a file:

I. Input the type of file F and read it.
II. Assign page number =1
III. Read the header H, footer F1, document summary D
IV. "WE" is used to read paragraphs of F
V. Print the details of F

4.2 Explicit Data Requests

In problems of class EF, the distributor is not allowed to add fake objects to the distributed data. So, the data allocation is fully defined by the agents' data requests. Therefore, there is nothing to optimize. In EF problems, objective values are initialized by agents' data requests [7].

Algorithm 1.
Agent Selection e-random

1: function SELECTAGENT (R, R1, . . ., Rn)

2: i ← select at random an agent from R

3: return i

Algorithm 2.
Agent Selection for e-optimal

1: function SELECTAGENT (R, R1, . . .,Rn)

2:i← max$\left(\frac{1}{R} - \frac{1}{R+1}\right)$

3: return i

Algorithm 2 makes a greedy choice by selecting the agent that will yield the greatest improvement in the sum objective. The cost of this greedy choice is $O(n^2)$ in every iteration.

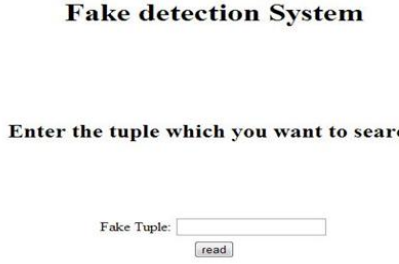
The overall running time of e-optimal is

$$O(n+n^2B)=O(n^2)$$

4.3 Sample Data Requests

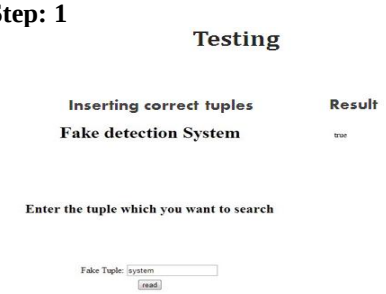
With sample data requests, each agent U may receive any T subset out of T(m) different ones. Hence, there are T(m) n=1 different object allocations. In every allocation, the distributor can permute T objects and keep the same chances of guilty agent detection [7]. The reason is that the guilt probability depends only on which agents have received the leaked objects and not on the identity of the leaked objects.

5. RESULTS



Testing Phase:

Step: 1



Step: 2

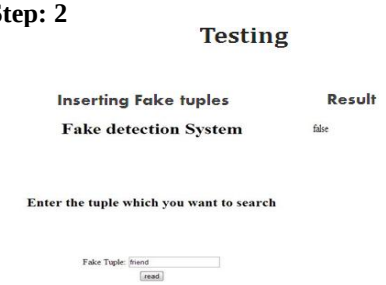


Fig: 4 Testing results

6. CONCLUSION

From above discussion we consider data leak as the most dangerous threat for organizations. An employee as an insider can purposely or fortuitously leak sensitive information. This sensitive information can be involuntarily distributed via Internet as e-mail, instantaneous messaging, spreadsheets, databases, and any other electronic means available – all without your awareness. To appraise the risk of distributing data two things are important, viz.,

(i) is data distribution strategy that helps to distribute the tuples among clients with least overlap and

(ii) Is calculating guilt probability which is based on overlapping of his data set with the leaked data set?

In spite of these troubles, we have shown that it is probable to evaluate the chances that a mediator 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 [4]. Our model is comparatively straightforward, but we consider that it confines the elementary requirements. The algorithms we have presented execute a variety of data allotment approach that can progress the distributor chances to identify a Traitor. We have shown that distributing objects shrewdly can make a momentous difference in identifying culpable agents, especially in cases where there is large have common characteristics in the data that agents must receive. A preface discussion of such a copy is available in [1] [2]. Now, industry & various offices can rely on this security & detection model.

7. FUTURE SCOPE

As per my concern the developer of an application can never be carried out to the fullest extend in a fixed time, the main reason why revisions of the application are always introduced in course of time. This application being restricted to one time development will have no amendment done, hence certain areas that can be improved

is pointed. The improved version of this system can detect the blameworthy agent even through the internet (online).

8. ACKNOWLEDGMENTS

We thank our H.O.D Prof. Dr.N. Chandra Sekhar Reddy for giving us the eminent facilities to perform my paper work. I am obliged to my faculty who has guided me all through the work.

9. REFERENCES

[1] IEEE Trans. on knowledge and data engineering, vol. 23, no. 1, January 2011.P. Papadimitriou and H. Garcia-Molina, "Data Leakage Detection".

[2] "Unauthorized Data Access Detection by Insertion of Duplicate Data Records". International Journal of Research in Computer and communication technology, IJRCCT, ISSN 2278-5841, Vol 1, Issue 7, December 2012.Y. Jagadeesh Kumar, M. Chinna Rao.

[3] International Journal of Emerging Trends in Engineering and Development Issue 3, Vol.1 (January 2013)http://www.rspublication.com/ijeted/ijeted_index.htm[1] Shivappa M Met agar [2]Theja N [3]Jayaprasad Motupalli.

[4] International Journal of Engineering Research and Development. "Detection and Avoidance of Data Leakage" Sandesh D Manocharya[1], A. Prabhakar[2]

[5] A model for improving guilt probabilities to Identify data leakages [1]A.Mounika, [2]M. Satwik, [3] G. Rajesh Chandra

[6] "Provenance in Databases,"Proc. ACM SIGMOD, pp. 1171-1173, 2007. P. Buneman and W.-C. Tan

[7] International Journal of Advanced Research in Computer Engineering & Technology Volume 1, Issue 4, June 2012. "Evaluate fake objects to predicts Unauthenticated developers".

[8] International Journal of Advanced Computer and Mathematical Sciences. ISSN 2230-9624 Vol 3, Issue 3, 2012, pp 337-342 http://bipublication.com D. Krishna Madhuri, A. Jagadeswara Rao, K. C. Ravi Kumar.

[9] International Journal of Computer Trends and Technology- volume3Issue4- 2012. "Development of Data leakage Detection Using Data Allocation Strategies".