ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 4, April 2013

DATA LEAKAGE DETECTION IN NETWORKS

Ms. Aishwarya Potdar[1], Ms. Rutuja Phalke[2], Ms. Monica Adsul[3], Ms.Prachi Gholap[4]

B.E, Department of Computer Engineering, KJCOEMR, Pune University, Pune, India[1,2,3,4]

Abstract— In the field of business, the owner of any organisation, company or business firm having some crucial data may need to share it with third-parties. These trusted third-parties may use this data for their own benefit causing reputational and monetary damage to the owner's company. If some of the shared data is discovered at some illegal place, it is quite possible that one or more third party agent is responsible for such information leakage. The owner must identify the leakage at the earliest and possibly the source of leakage. Data leakage is a silent type of threat. This sensitive data can be electronically transferred via e-mail, Web sites, FTP, instant messaging, spreadsheets, databases, and any other electronic means available – all without knowledge of the owner. Data allocation strategies (across the agents) are proposed that improve the probability of identifying leakages.

Keywords- Distributor, Allocation Strategies, Fake records, Guilty Agents, Request types.

I. INTRODUCTION

In today's technically empowered data rich environment, it is a challenge for data holders to prevent the leakage of data. Loss of large volumes of confidential information has become regular headline event, which force the companies to re-issue cards, inform customers and mitigate loss of goodwill from negative publicity.

While considering the protection of company's electronic assets from outsider threats – from intrusion prevention systems to firewalls to vulnerability management – organizations now turn their attention to an equally dangerous situation: the problem of data disclosure by the insiders. Whether its email, instant messaging, webmail, a form of website, or a file transfer, electronic communications exiting the company still go largely uncontrolled and unmonitored on their way to their destinations with the ever present potential for confidential information to fall into wrong hands.

II. EXISTING SYSTEM

Watermarking:

Nowadays, the digital assets such as software, images, video, audio and text are pirated which is a strong concern for owners of these assets. The protection schemes for such assets are based upon the insertion of digital watermarks into them. Watermarking means a unique code is embedded in each distributed copy. Data leakages can be identified using these embedded codes. Thus watermarking is a useful methodology. But in some cases the watermarks can be destroyed if the data recipient is malicious. Hence this technique proves to be inefficient.

III. PROPOSED SYSTEM

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. Fake objects are objects generated by the distributor that are not in the original set. The objects are designed which appear realistic, and are distributed among the agents along with the original objects. Different fake objects may be added to the data sets of different agents in order to increase the chances of detecting agents that leak data.

IV. PROBLEM SETUP AND NOTATIONS

T=Main Data Set
T={t1,t2,…….,tm}
U=Agent
U1, U2, U3,….., Un=Agent Set
Ri=Request
The agent can make a request to the distributor in two ways either sample request or explicit request.

- Sample request Ri=SAMPLE(T,mi): Any subset of mi records from T can be given to Ui

Example:

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 4, April 2013

Say that T contains patient records for a given hospital. If agent U_1 requests for any random 100 records of patients then such a request is called as sample request.

- Explicit request R_i=EXPLICIT(T,cond$_i$):

Agent U_i receives all T objects that satisfy cond$_i$.

Example:

In a hospital system, if an agent makes a request for records of patients satisfying a particular condition then such a request is explicit request.

- *System Diagram:*

Following figure shows the system diagram for describing the concepts of data leakage detection.

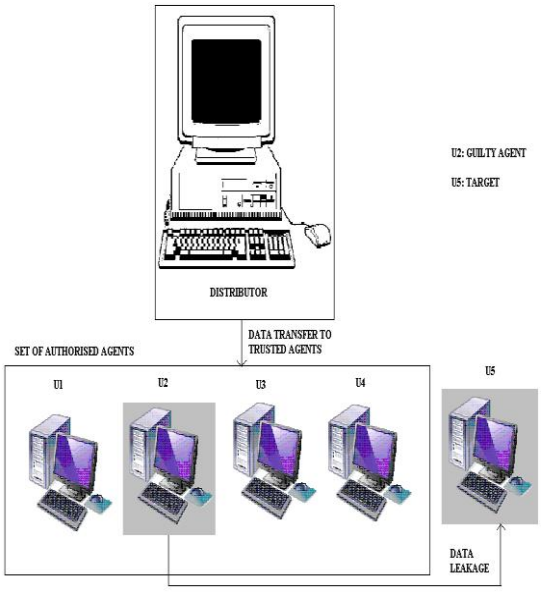

Fig. 1. System Diagram

The distributor is the owner of the data.U_1,U_2,U_3,U_4 are the supposedly trusted agents .

The distributor can send data to these agents by inserting different fake objects into the data sets of different agents. Now, suppose the distributor discovers his sensitive data at an unauthorised party U_5.The fake object present in the data set of U_5 can help the distributor to identify which of the four agents has leaked his data. Now, here agent U_2 leaked the sensitive data to U_5, then U_2 is known as the Guilty Agent and U5 is known as Target.

- *Distributor:*

The distributor is the main owner of the data.

- *Agents:*

These are supposedly trusted third parties who can make requests for data to the distributor.

- *Guilty Agent:*

The agent who leaks the sensitive data of the distributor to unauthorised party.

- *Target:*

The unauthorised party who receives the distributor's sensitive data leaked by the guilty agent.

V. MODULE DESCRIPTION

- *Database Maintenance:*

The sensitive data which is to be handed over to the agents is stored in the database.

- *Agent Maintenance:*

The registration detail about the agents as well as the data which is given to them by the distributor is maintained.

- *Addition of Fake Objects:*

The distributor is able to add fake objects in order to improve the effectiveness in detecting the guilty agent.

- *Data Allocation:*

In this module, the original records fetched according to the agent's request are combined with the fake records generated by the administrator.

- *Calculation Of Probability:*

In this module, the request of every agent is evaluated and probability of each agent being guilty is calculated.

VI. SYSTEM FLOW

A distributor can insert original as well as fake records in the Database. A new agent can be registered by entering personal details. A registered agent can Login and make a request to the distributor for data. The request can be of two types- Sample or Explicit.

The system then extracts the requested data from the main database and performs the addition of fake records to the set of original records. It then provides this data to the agent. The agent may pass on this data to an unauthorised party.

The agent or the unauthorised party may leak the sensitive data on the internet, television or other media via email, instant messaging, webmail or by any other means.

Whenever the distributor discovers the leaked set of his data, he will note the objects present in it. He will compare these objects to those present in the data sets handed over to different agents. He will list out the names of the agents whose data set contains the objects found in the leaked set. The fake records present in the data sets of the agents will be checked and the probability of an agent being guilty will be computed for each agent in the list.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 4, April 2013

The agent having the maximum probability will be the guilty agent.

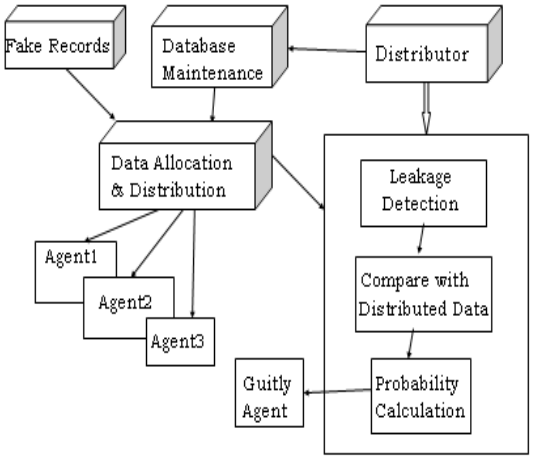

Fig. 2. Architectural View of the System

VII. CONCLUSION

This paper provides allocation strategies including the insertion of fake objects which will help identify data leakages and the guilty parties. Scope of this system can be extended by making provisions for the generation of fake records dynamically according to the agent's request.

ACKNOWLEDGEMENT

We sincerely thank Prof. V.B. Maral,our project guide, Prof. S.A Hirve, our project co-ordinator, Prof. S.S. Das, Head of the Department, KJCOEMR, Pune and Dr.S.J Wagh, Principal, KJCOEMR, Pune for their constant encouragement and motivation to write this paper.

REFERENCES

[1] Data Leakage Detection, *Panagiotis Papadimitriou*, Hector Garcia - Molina (2010), IEEE Transactions on Knowledge and Data Engineering, Vol 22,No 3

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

[3] S. Czerwinski, R. Fromm, and T. Hodes, "*Digital Music Distribution and Audio Watermarking*," http://www.scientificcommons. org/43025658, 2007.

[4] Papadimitriou P, Garcia-Molina H. A Model For *Data Leakage Detection*// IEEE Transaction On Knowledge And Data EngineeringJan.2011.

BIOGRAPHY

Ms.Aishwarya Potdar
born in January1992 in Pune, currently pursuing Computer Engineering in KJCOEMR, Pune.

Ms.Rutuja Phalke
born in September 1991 in Dhule, currently pursuing Computer Engineering in KJCOEMR, Pune.

Ms.Monica Adsul
born in December 1991 in Pune, currently pursuing Computer Engineering in KJCOEMR, Pune.

Ms.Prachi Gholap
born in November 1989 in Pune, currently pursuing Computer Engineering in KJCOEMR, Pune.