International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 3 536 - 539

Data Leakage Detection using Fake Data for Identifying Guilty Agents

Mr.Lukesh Kadu
Assistant Professor,
Shah and Anchor Kutcchhi
Engineering College.
Mumbai, India.
Email:sakec.lukeshk@gmail.com

Amol Gawas
Student at Shah and
Anchor Kutcchhi
Engineering College.
Mumbai, India.
Email:amolgawasdba@gmail.com

Prashant Verma
Student at Shah and
Anchor Kutcchhi
Engineering
College.
Mumbai, India.
Email:pv0793@gmail.com

Kishan Vaghani
Student at Shah and
Anchor Kutcchhi
Engineering College.
Mumbai, India.
Email:vaghanikishan2910@gmail.com

Chanda Sharma
Student at Shah and
Anchor Kutcchhi
Engineering College.
Mumbai, India.
Email:chanda.131993@gmail.com

***Abstract*—** The Data distributor allocates the confidential Data to the trusted agents for different work purposes. But sometimes the agent leaks the data to some other party. So in order to stop this leakage and identify the guilty agent we add fake data along with the original data using different data allocation algorithm. In this way Guilty agent can be identified and data leakage can be stopped.

Keywords*-Data leakage,Data privacy,Fake records,Guilty model,Data Allocation*

*****

I. INTRODUCTION

Now days, every organization is facing data leakage. That is very serious problem faced by organization. Data Leakage is the unauthorized transmission of private or sensitive data or information from within an organization to a third party [1].In the real world scenario, a distributer needs to share sensitive data among various stakeholders such as employees, business partners and customers. This increases the risk that confidential information will fall into unauthorized hands. For example, a hospital may give patient records to researchers who will devise new treatments. Similarly, a company may have partnerships with other companies that require sharing customer data. Another enterprise may outsource its data processing, so data must be given to various other companies [6]. We call the owner of the data the distributor and the supposedly trusted third parties the agents. Our goal is to detect when the distributor's sensitive data have been leaked by agents, and if possible to identify the agent that leaked the data [2]. Traditionally, leakage detection is handled by watermarking, e.g., a unique code is embedded in each distributed copy. Water-marks can be very useful in some cases, but again, involve some modification of the original data. We refer 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. In our system, we develop a model for assessing the guilt of agents. We also present algorithms for distributing objects to agents, in a way that improves our chances of identifying a leaker. Finally, we also consider the option of adding fake objects to the distributed set . Such objects do not correspond to real entities but appear realistic to the agents. In a sense, the fake objects act as a type of watermark for the entire set, without modifying any individual members. If it turns out that an agent was given one or more fake objects that were leaked, then the distributor can be more con dent that agent was guilty[3].

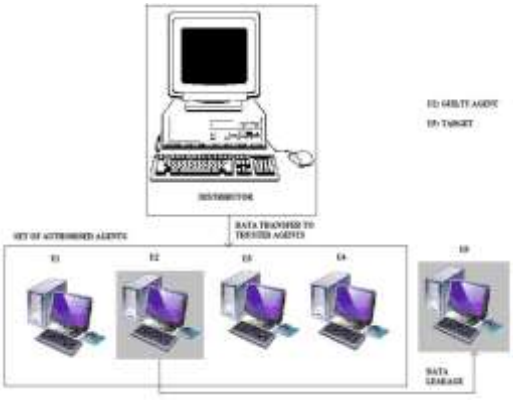

Fig. 1 Data Leakage Example

II. OBJECTIVE

1. The objective is to make sure that the agent who leaks the data is identified.
2. To avoid modifying the original data while allocating since data originality matters and could be sensitive in some cases.

3. To make use of best algorithms while allocating the data to the agents.
4. To make sure that the purpose for which the data are given to the agents are achieved.
5. To avoid manual allocation of data that is merging of real and fake data.
6. Once the agent is caught it will be made sure that the data are not given to that agent again and he/she that is the guilty agent will be blacklisted.

III. EXISTING SYSTEM

The most common technique that is watermark technique, is used in data leakage. In Watermark technique a uniquely identified text or image is embedded within each copy that is allocated to authorized agents. When leakage occurred, then this unique code would help identify the party that was responsible for the leak. The problem with this approach was that even though this is an easy solution, it still involves a certain modification of the original data information set. Also, it was observed that such watermarks could be tampered with to sufficiently distort the uniquely identifying code or sometimes completely destroyed if the data recipient is malicious. Also it is very time consuming and also distributor can claim the ownership of the data since he/she modifies it.[6]

IV. PROPOSED SYSTEM

In this, we propose to add the fake data while allocating the data to the agents rather than using the watermark technique. By adding the fake data there will be no need to modify the original data. The data merging will be done using different algorithms depending on the types of requests. Once the data leakage is detected using fake data. The data given to the agents will be compared and the probability will be calculated and based on this the guilty agent will be identified.

MODULE

1) Database maintenance:
Here the agent registration details are maintained and the sensitive data which are provided to agents are specified. The designing of the whole database is done.

2) Agent maintenance:
a. Registration
Here details of agents are registered and it collects the information about them like what are the sensitive data they want.

b. History
Here the agent history is maintained like what all the details are given by distributor previously. It maintains entire details of the agent. To detect the guilty agents it checks the history and detects those agents who have fake details from the third party.

c. Blacklisted Agent
In this the records are maintained of the Agents who had leaked the data in past. This helps distributor to see that he is not giving the data to the Agent who has leaked some data in the past.

3) Detecting Guilty Agent:
The Distributor will allocate the data to different agents using different allocation strategy method depending on the type of data request made by the agent. If the data which has been allocated to the agents if found in some other domain, then the distributor can compare the data with fake that was allocated to the agents based on which the probability will be calculated. The probability that is calculated then will be compared and whose probability will be more that agent will be the guilty agent.

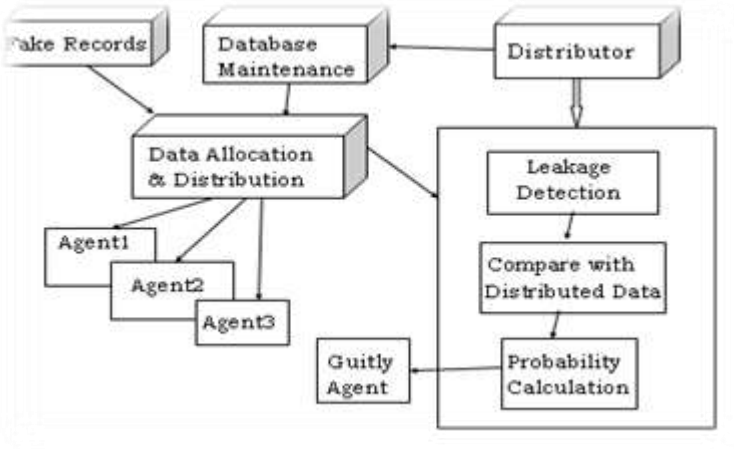

Fig. 2 System Architecture

V. DATA ALLOCATION

In Data allocation, Data are allocated to the agent by the distributor. The agent makes the request to the distributor. The request made by the agent could be implicit data request or explicit data request. In implicit data request agent would just ask for data with no condition, whereas in explicit data request the agent will give certain condition for example He/she might need data of people between ages 18 to 40.

Basic Data Allocation:
In this the distributor will allocate Data to the Agent with the help of the basic algorithm 1, where original and fake data will be merged and will be given to the agent. The ratio of fake data and original data will be 1:2.In this algorithm 1 the table will be created where the real records will put in along with the fake data. The limit of the record to be given will be checked and would allocate data accordingly.

Algorithm 1:

Input: AGENT, filename f{x}, record{O}, AGENT, filename f{x}, record{O}, record{f}.

Output: Result R{}, f{x} //file

Step1: generate CREATE TABLE f{x}

Step2: INSERT INTO f{x} from record{O}

Step3: lim = (lim / 2) //predefined Ratio
(lim is no. of records agent asked for).

Step4: ORDER BY rand()

Step5: INSERT INTO f{x} from record{f}WHERE cond

Step6: ORDER BY rand()

Step7: o1{record(O)} + f1{record(f)} = R{};

Step8: R{} SELECT INTO OUTFILE AGENT/f{x}

Explicit Data Allocation:
In this the Distributor will allocate the data to the agent for the explicit data request with the help of the algorithm 2.In this algorithm 2 table will be created first then the data will be allocated as per the condition of the agent in which fake data will be added randomly and then it would be given to the agent.

Algorithm 2:

Input: AGENT, filename f{x},lim, //number of records
record{O}, record{f},cond.

Output: R{}, f{x} //file

Step1: generate CREATE TABLE f{x}

Step2: INSERT INTO f{x} from record{O}WHERE cond

Step3: lim = (lim / 2) //predefined ratio

Step4: ORDER BY rand()

Step5: INSERT INTO f{x} from record{f} WHERE cond

Step6: ORDER BY rand()

Step7: o1{record(O)} + f1{record(f)} = R{};

Step8: R{} SELECT INTO OUTFILE AGENT/f{x}

Linear Congruential Random Number Based Algorithm:
With the help of this method merging of real data and fake data is done more conveniently where it does not follow a particular ratio of merging real data with the fake. It merges data randomly does reducing the overlapping of data and identifying the guilty agent becomes easier. Also distinguishing between fake and original data is not possible due to reduction in data overlapping.

Formula -> Xn+1 =(A*Xn + B)mod m
where A,B,M are prime numbers & Xn th term is random number
NOTE: random number generated should be greater than half of lim
(lim is no. of records agent asked for).

If the agent asks for suppose lim=10 records, split 10 into two halves such that any one halve is greater than 5 and less than 9.

New_lim = lim - Genrated_Random_no.

Now we have to split lim into two.
The greater of two will be limit for clean data and other one for fake data.
Lets say for this algorithm greater of
(new_lim,Genrated_Random_no.) is new_lim

Algorithm :

Step1: generate CREATE TABLE f{x}

Step2: INSERT INTO f{x} from record{O}WHERE cond LIMIT new_lim

Step3: ORDER BY rand()

Step4: INSERT INTO f{x} from record{f} WHERE cond LIIMIT Genrated_Random_no.

Step5: ORDER BY rand()

Step6: o1{record(O)} + f1{record(f)} = R{};

Step7: R{} SELECT INTO OUTFILE AGENT/f{x}

VI. CONCLUSION

The paper addresses a different problem where data transmission takes place through human beings known as trusted agents. Detecting the Guilty agent is the most important aim of this paper, since the agents are leaking the confidential data. The identification of the guilty agent is a challenging task. So in the process of data leakage detection we came across the watermark technique which is not sufficient, since in this technique data is modified and also watermark code can be easily tampered. So adding of fake data along with the original data is decided. The data allocation is done through different algorithms depending on type of data requests. And once the Data leakage is detected, the distributed data among the agents will be compared and the probability will be calculated and the guilty agent will be identified and it will be blacklisted and made sure that the agent is not given any data in future for wok purpose. But to avoid data overlaps and to make data merging more useful Linear Congruential Random Number Based Algorithm should be used instead of basic data allocation so as to identify guilty agent easily and also agent cannot distinguish between fake and original data with this method since data overlapping will be reduced.

VII. REFERENCES

[1] Panagiotis Papadimitriou, Member, IEEE, Hector Garcia-Molina, Member, IEEE., *Data Leakage Detection*, IEEE Transactions On Knowledge And Data Engineering, Vol. 23, No. 1, January 2011
[2] P. Papadimitriou and H. Garcia-Molina, *Data Leakage Detection* , technical report, Stanford Univ., 2008
[3] Keerthi.P,M.Sheshikala,D.Rajeswara Rao,*Guilty Agent Detection by Using Fake Object Allocation*", International Journal of Computer Technology Volume -1,2013
[4] Rudragouda G Patil, *Development of Data Leakage Detection Using Data Allocation Strategy* International

Journal of Computer Applications in Engineering Sciences,VOL I, Issue II, June 2011

[5] Ajay Kumar ,Ankit Goyal ,Ashwani Kumar ,Navneet Kumar Chaudhary ,Sowmya Kamath S ,"Comparative Evaluation of Algorithms for Effective Data Leakage Detection", Proceedings of 2013 IEEE Conference on Information and Communication Technologies *(ICT)*, 2013.

[6] Ahirrao P. P., Rai S. S., Pathania B. R., "Data Leakage Detection", International Journal of Recent Technology and Engineering (*IJRTE*) ISSN: 2277-3878, Volume-3, Issue-1, March 2014

[7] Archana Vaidya, Prakash Lahange, Kiran More, Shefali Kachroo & Nivedita Pandey, Data Leakage Detection, International Journal of Advances in Engineering & Technology, March 2012

[8] Priya Walunj, Priya Tadge, Navnath Kondalkar, Satish Mahamare, Identification of Data Leakage and Detecting Guilty Agents Using Data Watcher, International Journal of Advanced Research in Computer Science and Software Engineering,February 2015