Volume 3, No. 6, June 2012
Journal of Global Research in Computer Science
ISSN-2229-371X
RESEARCH PAPER
Available Online at www.jgrcs.info

AN EFFICIENT AND ROBUST MODEL FOR DATA LEAKAGE DETECTION SYSTEM

Janga Ajay Kumar [1] and K. Rajani Devi [2]
[1]Student, M.Tech (IT), Siddharth Nagar, Nalanda Institute of Engineering and Technology, A.P, India.
[1]ajaykumarjanga002@gmail.com
[2]Head of the Department, (IT), Siddharth Nagar, Nalanda Institute of Engineering and Technology, A.P, India.

Abstract- In every enterprise, data leakage is very serious problem faced by it. An owner of enterprise has given sensitive data to itsemployee but in most of the situation employee leak the data. That leak data found in unauthorized place such as on the web of comparator enterprise or on laptop of employee of comparator enterprise or the owner of comparators laptop. It is either observed or sometimes not observed by owner. Leak data may be source code or design specifications, price lists, intellectual property and copy rights data, trade secrets, forecasts and budgets. In this case the data leaked out it leaves the company goes in unprotected the influence of the corporation. This uncontrolled data leakage puts business in a backward position. To find the solution on this problem we develop two models. First, when any employee of enterprise access sensitive data without the consent of owner in that case ,we developed data watcher model to identifying data leaker and suppose employee given data outside the enterprise for that we devolved second model for assessing the "guilt" of agents. Guilt model are used to improve the probability of identifying guilty third parties.For implementing this system, we used SSBT'S COET, Bambhori, and Jalgaon college database. In this system we consider, data owner is college management called as distributor and other employee is called as agents. For that we considered two condition sample or explicit condition because agents want data in sample or condition.

Keywords- Sensitive data, Fake Data, DataRequest, Guilt Model.

INTRODUCTION

Inenterprise, owner must hand over sensitive data to supposedly trusted agents For example; financial data give to the financialemployee for making balance sheet or for making financial transaction but that data was leaked out. Similarly, a company may have partnerships with other companies that require sharing customer data. 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 [1]. However, in some cases, it is important not to alter the original distributor's data. For example, if financial data cannot be perturbation. If medical researchers will wants exact data of patients. They may need accurate data for the patients. Traditionally, leakage detection is handled by watermarking e.g., a unique code is embedded in each distributed copy.

If that copy is later discovered in the hands of an illegal party, the leaker can be identified. Watermarks can be very useful in some cases, but again, involve some modification of the original data. In addition, watermarks can sometimes be cracked if the data recipient is malicious. In this paper, we study unobtrusive techniques for detecting leakage of a set of objects or records [7][8].

Specifically we study the following scenario: In everyenterprise, data leakage is very serious problem faced by it. An owner of enterprise has given sensitive data to its employee but in most of the situation employee leak the data. That leak data found in un- authorized place such as on the web of comparator enterprise or on laptop of employee of comparator enterprise or the owner of comparators laptop. It is either observed or sometimes not observed by owner. Leak data may be source code or design specifications, price lists, intellectual property and copy rights data, trade secrets, forecasts and budgets.

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. If the distributor sees "enough proof" that an agent leaked data, he may stop doing business with him, or may initiate legal proceedings. In this paper, we develop a model for assessing the "guilt" of agents. Such objects do not correspond to real entities but appear practical 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 surer that agent was guilty [1].

OBJECTIVE

A data infringe is the inadvertent release of secure information to a not trusted environment. The goal is to estimate the likelihood that the leaked data came from the agents as opposed to other sources. Not only to we want to estimate the likelihood the agents leaked data, but we would also like to find out if one of them in particular was more likely to be the leaker with large number of overlapping. The data allocation strategies help the distributor "cleverly" give data to agents. Fake objects are added to identify the guilty part, to address this problem four instances are specified. Depending on which the data request is provided. Depending upon the type of data request, the fake objects are allowed.

A distributor owns a set T = {t1, t2, t3…tm} of valuable data objects. The distributor wants to share some of the objects with a set of agents U1, U2….Un, but does not wish the objects be leaked to Other third parties. The objects in T could be of any type and size, e.g., they could be tuples in a relation, or relations in a database.

An agent Ui receives a subset of objects Ri $\subseteq$ T, determined either by a sample request or an explicit request: Sample request Ri = SAMPLE (T, mi): Any subset of mi records from T can be given to Ui.

Explicit request Ri = EXPLICIT (T, Condi): Agent Ui receives all the T objects that satisfy Condition [1].

BACKGROUNDAND MOTIVATION

Data Leakage Worldwide Common Risks and also the Mistakes Employees Make examined the whole a relationships between employee behaviorsand data loss, as well as IT perceptions of those factors. Then a survey found that employees around the world are engaging in behaviors that put corporate and personal data at risk that IT professionals are often unaware of those behaviors, and that preventing data leakage is a business-wide challenge [2].

The helpfulness of security policies, offered insight into how security policy creation, communication and compliance effect data leakage. The analysis showed that a lack of security policies and a lack of employee compliance with security policies were significant factors in data loss. And as in the first set of findings, the survey showed that IT professionals lacked important awareness-in this case about how many employees actually understand and observe with security policies. Thus it is concluded that companies must address the dual challenge of creating security policies and enforcing employee compliance [2].

The guilt detection approach is related to the data provenance problem [9] tracing the lineage of S objects implies essentially the detection of the guilty agents. Tutorial [3] provides a good over- view on the research conducted in this field. Suggested solutions are domain specific, such as lineage tracing for data warehouses [4], and assume some prior knowledge on the way a data view is created out of data sources. Our problem formulation with objects and sets is more general and simplifies lineage tracing. As far as the data allocation strategies are concerned, our work is mostly relevant to watermarking that is used as a means of establishing original ownership of distributed objects. Watermarks were initially used in images [5], video [6]. Watermark cannot be inserted. In such cases, methods that attach watermarks to the distributed data are not applicable. Finally, there are also lots of other works on mechanisms that allow only authorized users to access sensitive data through access control policies. Such approaches pre- vent in some sense data leakage by sharing information only with trusted parties [5].

An employee who is disgruntled or seeks to gain profit through illegal actions that involve corporate resources can become an insider threat that adds a dangerous new dimension to the data loss prevention challenge. The disgruntled insider threat defines a common awareness that the most significant security threats originate outside the company. Employees with a spiteful agenda and a profit motive can use their insider status to engage in activities that cause even greater financial loss than external threats. Rightful network access and stewardship of devices such as laptops and PDAs makes it simple for disloyal employees to leak corporate data [2].

Some employees simply fail to return company devices when they leave a job.

This is an expensive and dangerous activity for businesses because it adds yet another avenue for data loss. Even if only 5 percent of exiting employees take a device, that adds up to50 employees in a company of 1000, or 500 in an enterprise of10,000 employees. For larger organizations, the financial and data loss risks are far more significant.

A shocking 11 % of employees reported that they or fellow employees accessed unauthorized information and sold it for profit, or stole computers. Employee reasons for keeping their corporate devices when leaving a job included needing the device for personal use (60 %), getting back at their companies, and a belief that their previous employers would not find out.20 % of IT professionals said disgruntled employees were their biggest concern in the insider threat arena [2].

EXISTING SYSTEM

In many cases distributor must indeed work with agents that may not be trusted, and distributor may not be sure that a leaked object came from an agent or from some other source, since sure data cannot admit watermarks. In existing system there is few problem like fixed agents and existing system work comparable with agents whose request known in advance. Also with adding fake object original sensitive data cannot be alter and absences of agent guilt models that capture leakage scenarios and appropriate model for cases where agents can collude and identify fake tuples. Lastly system is not online capture of leak scenario also in existing sys- tem more focus on data allocation problem.

PROPOSED SYSTEM

To find the solution on this problem we develop two models. First, when any employee of enterprise access sensitive data without the consent of owner in that case, we developed data watcher model to identifying data leaker in this point suppose data leaker will identify then no need to calculating the probability of agents that method gives near about 90 % of result. But suppose employee given data outside the enterprise for that we devolved second model for assessing the “guilt” of agents. Guilt model are used to improve the probability of identifying guilty third parties.

For implementing this system we used SSBT’S COET, Bambhori, and Jalgaon college database. In this system we consider data owner is college management called distributor and other employee is called agents. For that we take two condition sample or explicit condition because agents want data in sample or condition. In this approach, the model for assessing the “guilt” of agents is developed.

The option of adding "fake" objects to the distributed set is considered. Such objects do not

Correspond to real entities but appear practical to the agents. In a sense, the fake objects acts as a type of watermark for the entire set, without modifying any individual members. If it turns out an agent was given one or more fake objects that were leaked, then the distributor can be more confident that agent was guilty.
Proposed System worked on two processes

Data Distribution Process:

In that considered two exiting techniques for data allocating to the agents. There are four instances of this problem they address, depending on the type of data requests made by agents (E for Explicit and S for Sample requests) and whether "fake data" are allowed (F for the use of fake data, and F for the case where fake data are not allowed). Fake data are data generated by the distributor that are not in set T. The data are designed to look like real data, and are distributed to agents together with the T data, in order to increase the chances of detecting agents that leak data [1].

Probability Finding Process:

While distributing the data to any agents some kind of receiver's information can be added to find out the guilty agent it is more concentrated on finding the probability of an agent to be found as guilty. Data object is to be important aspect of our work; it is consider agents parameter and overlapping between pair of agents of this data object which we are forwarding to other agent. The parameter would then be checked once a data object is received from a malicious target for that used a special process for data object is received from any target the probability is calculated the data object came from which source or we can guess that which agent has leaked the data. Guilty Agent Model would be used to find the agent to be guilty with numerous conditions. Also we have considered if the object cannot be guessed or if its probability can't be find out then the agent can't be considered to be guilty.

Algorithm for Find Guilt Agent:

a. Distributor select agent to send data. Distributor selects the agents to send the data according to agent request.
b. Distributor creates fake data and allocates it to the agent. The distributor can create fake data and distribute with agent data or without fake data. Distributor is able to create more fake data;he could further improve the chance of finding guilt agent.
c. Check number of agents, who have already received data. Distributor checks the number of agents, who have already received data.
d. Check for remaining agents. Distributor chooses the remaining agents to send the data. Distributor can increase the number of possible allocations by adding fake data.
e. Estimate the probability value for guilt agent. To compute this probability, we need an estimate for the probability that valuescan be "guessed" by the target.

IMPLEMENTATION

In every enterprise, data leakage is very serious problem faced by it. An owner of enterprise has given sensitive data to its employee but in most of the situation employee leak the data. That leak data found in unauthorized place such as on the web of comparator enterprise or on laptop of employee of comparator enterprise or the owner of comparators laptop. It is either observed or some- times not observed by owner. Leak data may be source code or design specifications, price lists, intellectual property and copy rights data, trade secrets, forecasts and budgets. In this case the data leaked out it leaves the company goes in unprotected the influence of the corporation. This uncontrolled data leakage puts business in a backward position. Suppose employee given data outside the enterprise for that we devolved second model for assessing the "guilt" of agents. Guilt model are used to improve the probability of identifying guilty third parties.

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 [1][7].

In the figure Fig. (1) Distributor has been given the data to agents according to the request by agents. In this system database maintained by Distributor according to the request by agents data passing to agents with fake or without fake object. When agent doing the business with target without the consent of distributor and leak data. Distributor discovers some of those same objects in an unauthorized place. (For example, the data may be found on a web site, or may be obtained through a legal discovery process or someone's laptop). Then distributor match leak data with his data. Distributor also check overlapping of data among agents and then he calculating probability of agents varies for {0, 1}. Like this probability of agents varied on both ways differently apparently in this feature will be very useful on now a days to this approach we will make it as a agent guilt probabilities. Information will be retrieved by so many external experiments.

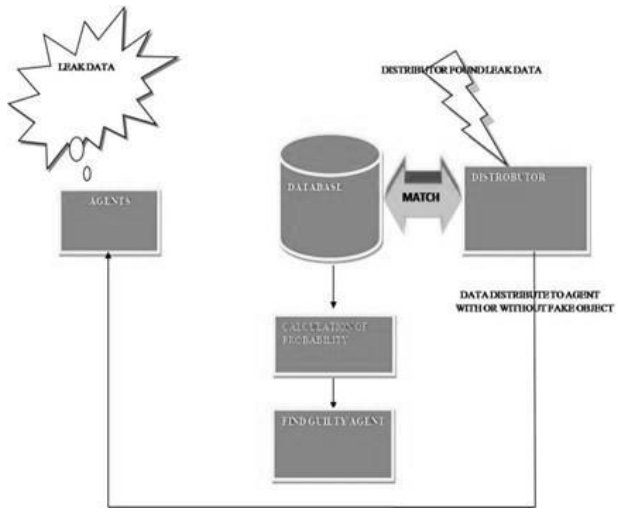

Figure. 1- Architecture of data leakage detection

RESULTS AND DISCUSSION

The effectiveness of a system is most commonly described with its "Record wise leak report" and "Probability of agent guilty".

$$\text{Record wise leak report} = \frac{|T \cap S|}{|S|}$$

This formula calculates records wise leak data without considering agents overlapping. According to that the distributor knows which agents consider for calculating guilt probability.

$$\text{Probability of agent guilty} = \frac{|R_i \cap S|}{|S|}$$

This formula calculates records wise leak data with considering agents overlapping mean that particular record share by how many agents. According to that the distributor knows which agents consider for calculating guilt probability. It will enhance the Agent Probabilities. Very Useful Probabilities of guilty are listed below.

agentGuiltyProbability

agentName	ProbabiltyOfguilty	Add New Field
Agent1	0.90496	
Agent2	0.266666666666667	
Agent3	0.784	
Agent5	0.56	
*	0	

Figure. 2 - Agent guilt probabilities

AgentProbabilityDistribution

agentName	valueOfP	Prob
Agent1	0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1	0.064,0.006,0.001,0,0,0,0,0,0,0
Agent2	0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1	0.7,0.513,0.394,0.315,0.262,0.227,0.205,0.191,0.1
Agent3	0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1	0.166,0.036,0.01,0.003,0.001,0.001,0,0,0,0
Agent5	0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1	0.385,0.169,0.084,0.047,0.03,0.02,0.016,0.013,0.0
*		

Figure. 3 - Agent guilt probability distribution

As shown In Fig.(2) given guilt probability of agents like agent1 having 0.9 probability means he having more probable to leak the owner data then agent3 having 0.7 then agent 5 having 0,5 like- wise owner can find out leak scenario. As shown in Fig. (3) First owner set estimate probability from 0.1 to 0.9 according to owner can find guilt agents as shown in Fig. (4), according to the record distribution to agents plot the graph in that considered probability value from {0 to 1} and estimate value according to assumption 2. Generally this thing will be set by owner of enterprise according to the liability of agent owner set estimate probability for 0.1 to 0.9.

Table 1 - Comparative analysis between existing system and proposed system

Algorithm Parameter	Existing Data guilt model	Data guilt model with data watcher
Techniques used	Checking overlap-ping between agents	Checking overlapping between agents
Creation of fake data	In every iteration	Depend upon the data
Condition used for allo-cating object to agent	Sample and explicit	Sample and explicit
Number of agents	2 or 3	More than 5 agents
Agent request	Known in advance	Not need due to data share watcher

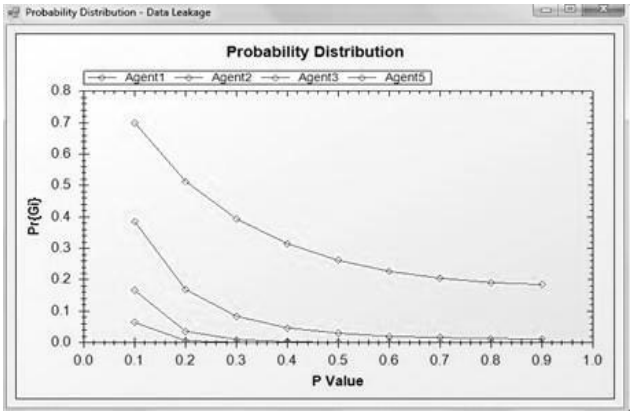

Figure.4- Graphical representation of probability Distribution

CONCLUSION AND FUTURE WORK

In the enterprise owner hand over its sensitive data to employee but before that owner must be add water mark to each and every sensitive data. Also check the history of employee means that particular employee is liable or not to handle that data or not. Then hand over data to employee. Suppose that employee leak data accidently for this case owner considering the estimate probability of employee. If employee leak data deliberately then owner think about that particular employee for not involved in shared data or confidential work or talk.

In malice of these difficulties, we have shown 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. The data distribution strategies improve the distributor's chances of identifying a leaker. It has been 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. In some cases "practical but fake" data records are injected to improve the chances of detecting leakage and identifying the guilty party. Our future work includes extension of this work considering allocation strategies so that they can handle agent requests uniquely in an online fashion and fake data using encryption techniques.

REFERENCES

[1]. Panagiotis Papadimitriou and Hector Garcia-Molina (2011) IEEE Transactions on Knowledge and Data Engineering, 23(1), 51-63.

[2]. Cisco white paper (2008) Data Leakage Worldwide: The High Cost of Insider Threats, 01-06.

[3]. Cui Y., Widom J. (2003) VLDB Journal, 12, 41-58.

[4]. Buneman P., Tan W.C. (2007) ACM SIGMOD international conference on Management of data, 1171-1173.

[5]. Ruanaidh J.J.K.O., Dowling W.J., Boland F.M. (1996) IEEE Vision, Signal and Image Processing, 143 (4), 250-256.

[6]. Oztan Harmanci, Kivanc Mihcak M., Murat Tekalp A. (2007) ICASSP, I-833-I-836.

[7]. Panagiotis Papadimitriou, Hector Garcia-Molina (2009) IEEE International Conference on Data Engineering. 1307-1310.

[8]. Agrawal R. and Kiernan J. (2002) 28th Int. Conf. Very Large Data Bases, 155-166.

[9]. Buneman P. and Tan W.C. (2009) ACM SIGMOD, 38 (2), 42-49.