RESEARCH PAPER | Engineering | Volume : 5 | Issue : 6 | June 2015 | ISSN - 2249-555X

Information Leak Discovery From Guilty Agent

KEYWORDS: Data allocation strategy, Fake objects, Guilty agents.

Savitri Vasudev Yadav
MCA Final Year, Student, Department of computer science, Amrita school of arts and sciences Mysore, India.

Barkath Fathima
MCA Final Year, Student, Department of computer science, Amrita school of arts and sciences Mysore, India

ABSTRACT *Data allocation technique is an emerging technology which carries potential to resolve the leakage of data in an integrated way. However, while doing corporate some of the sensitive data should be handover to some company. Some of the information is leaked and found in an unauthorised space. So distributor has to find the leak data. Presently watermarking technique is being used for the data security. But this technique doesn't provide sufficient protection. This report includes data allocation technique and fake objects that improve our opportunities to notice the leakage and recognize the guilty agent.*

I. Introduction

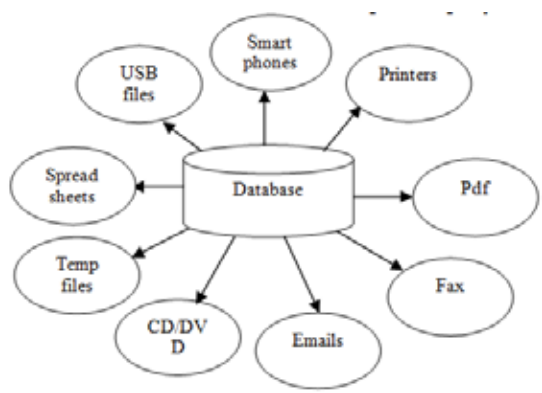
Information leakage is a problem in many institutes, organization of all countries. In business applications some of the sensitive data should be handover to some company to the trusted third parties. Sensitive data with any company which consist of intellectual property, financial information, personal credit card data and other information depending on the business. Leakage might be through mails, misbehaving the mail-ids, or hacking the mails or extracting the user's and passwords. By securing those misleading, it's really important to receive the information. Paying off those problems is very necessity in this advanced technology. But sometimes hackers get easily all the data from trusted third parties or from unauthorized places, for instance laptops, mobiles, and computer, etc. For solving those problems we have chosen data security and information security. Thither is a versatile to solve these troubles. One is by applying an algorithm called data allocation strategy.

Existing system

In existing system they considered applications where the original sensitive data cannot be distracted. The disturbance is a useful technique where the data can be modified and can be arrived at "less sensitive" before that is handed to agents. However, in some cases, it is significant not only altering the original distributor's data. Commonly, leak detection is handled by watermarking, example a unique code is embedded in each assigned copy. If that copy is later set up in the workforce by an unauthorized party, the leak can be found away. Watermarks can be very useful in some instances, but again, it took some alteration of the original data. Furthermore, watermarks can consistently be put down if the data recipient is malicious.

Proposed system

In proposing a system after passing on a bunch of objects to agents, the distributor discovers some of those same objects in an unauthorized place. At this point the distributor can assess the likeliness that the leaked information came from one or more agents, as opposed to having been independently collected by other means. If the distributor sees "enough evidence" that an agent leaked data, he may stop doing business with him, or may commence legal proceedings. In this project we prepare a model for evaluating the "guilt" of agents. We also present algorithms for passing around objects to agents, in a manner that betters our prospects of identifying a leader. In the end, we also consider the choice of adding "fake" objects in the distributed set. Such objects do not equate to real entities, but appear. If it turns out an agent was applied one or more fake objects that were leaked, then the distributor can be more convinced that the agent was guilty.



Example of information leakage detection

Literature study

Paper 1:This section gives a brief literature review of the "Data Leakage Detection" technique used in this paper is perturbation [1] that is a very useful technique where data is modified and less sensitive before handing to the agents. For example, one can add random missing data to attributes, or any data can be replaced by some ranges. In some cases it is not important to alter some data , for example: if any data outsource is doing our payroll he must have unique customer ID number exact salary. But in this paper they have implemented different algorithm for the different processing first algorithm used for finding for entities and agents.

Algorithm for Entites and Agents

Input: {s1..... Sn}

{u1.....un}

Volume : 5 | Issue : 6 | June 2015 | ISSN - 2249-555X

{condi…….condin}

Output:{r1……rn}

Step 1: Distributor sends data for set T={s1…….sn} of valuable data objects.

Step 2: Then the distributor wants to share some of the objects with a set of agents u1…….. un. But he does not wish the objects be leaked to the trusted parties.

Step 3: The objects in T could be of any types.

Step 4: An agent, ui receive a subset of the objects ri subset not equal to either by input or output request.

Step 5: First request r1=request(t1,mi):any r1 subset of mi records from T can be given to ui.

Step 6: Second agent ui receives all the T objects which satisfies condition.

Algorithm for guilty agent
Step 1: Suppose that after giving objects to agents the distributor discovers that a s is subset equal to has been leaked.ie; some to third parties.

Step 2: Then found out who have been leaked the information.ie; by finding probability that agent ai is a guilty agent given as evidence S.

Paper 2:In this paper they developed a paper for "A model for data leakage detection". They have studied algorithms for distributing objects to agents, in a way that improves our chances of identifying the agents. In this algorithm they have considered the options of adding "fake" objects in the distributed set. Such that objects do not correspond to the agents. [2] Technique used were perturbed where in making data in a less sensitive manner. To solve this problem algorithm used is data allocation problem setup.

Algorithm for data allocation problem setup

Input:{EF……EFn} and {s1…….sn}

{b1……bn}

{condi…….condin}

Output:{w1……wn}

Step 1: In probability the distributor does not allow to add the fake objects to the distributed objects.

Step 2: EF problem objectives, values are initialized by agents.

Step 3: B= {b1, b2} only two agents with input data requests. Ie; {s1,s2} and {s1,s2}.

Step 4: but distributor cannot remove or alter the b1 or b2 because to decrease the or to overlap the r1∩r2.

Step 5: if creation objects (b=1) and both can receive one fake object (b1=b2=1). Where distributor can add fake objects either r1, r2 to increase the correspondent denominator of summation term

Step 6: then receiving fake object o(n) till o(n+b).hence running the algorithm is o(n+b).

Step 7: b≥ $\sum_{i=1}$ bi to minimize even term of the objective and also a summation by adding the maximum bi objects, then finally finding out the set of leaked data through set r1.

Paper 3: The paper reviewed is "Data Leakage Detection and E-mail Filtering ". [10] The important data has given or sensitive data send to trusted agents ie, in unauthorized parties. This increases the risk that confidential information will pitfall into unauthorized place. For decreasing this problem they used have implemented for satisfying the condition agents requests. Their goal is to detect when distributor's sensitive data has been leaked by agents,and if possible to identify the agent leaked thae data and what data has been leaked. Data can be leak through e-mails. So there is need to provide security to the data. For this purpose e-mail filtering concept has been implemented in this system. Even if any agents sends distributor sensitive data to unauthorized person via, e-mail, unauthorized person will not able to see download that e-mail. By providing them with the number of objects that satisfy their condition. The distributor may not deny serving an agent request and may not denied serving an agent request and may not be provide agents with different same versions of the objects. For this purpose their main objective has maximized the chances of detecting a guilty agent that leaks all his information objects.

Modules for Models for data leakage detection system were:-
Fake objects module

Data allocation strategies

E-mail filtering

Fake objects: Fake objects are discovered by distributor in order to increase the chances for detecting agents that leak data. Fake objects is a objects that which look ike real objects.

Every data distributor owns data to agents. But position and numbers will be different from eachother. Depending upon the number of records the number of fake objects will also differ so that it will be easy to detect the guilty agent.

Data allocation strategies: The distributor unknowingly gives data to agent to improve the chances of detecting guilty agents. Data allocation depends on the request done by the agent and whether system can add fake objects in it.

The agent request can be done in tow types:-

Sample and explicit.

Sample: Agent receives a subset of distributors data which required by agents.

Explicit: Special data request which satisfies a special condition is given to agent.

creates and adds fake objects to the information that he deals out to agents. Fake objects are objects generated by the distributor in order to increase the chances of de-

Volume : 5 | Issue : 6 | June 2015 | ISSN - 2249-555X

tecting agents that leak data. [5] The distributor may be to add fake objects to the distributed data in order to improve his effectiveness in detecting guilty agents. Our usage of fake objects is inspired by the role of "trace" records in mailing lists. In case we give the wrong secret key to download the file, the duplicate file is opened, and that fake detail also sends the mail. Ex: The fake object details will display.

E-mail Filtering: In this module the mail is sent to unauthorized users and authorized users. Unauthorized receives a mail, but the system detect that mail has been send to unauthorized user. Then System filters the data and block the content of the mail. If unknowingly unauthorized users download that mail,the mail doesnot contain original content of the mail and the download file of the size will be zero.

Optimization Module is the distributor's data allocation to agents has one constraint and one target. The agent's constraint is to satisfy distributor's 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. User can able to lock and unlock the files for secure.

Proposed Data allocation algorithm:

Inputs:

{R1,Rr2….Rn} (Setof distributor sends the data)

{F1,F2….. Fn} (Set of Fake objects)

{i1,i2……...in} (Sets to increment the data)

{k……kn} (Set to add additional parameters)

{E……..En} (Represent leaked data)

{a……...an} (Represent numbers of agents leaked)

{B……..Bn} (Represent blocking the information)

{P……...Pn} (Represent all sample request objects)
{Q……Qn} (Represent all explicit request objects)

1. The request is sent by the agent to the distributor .
2. The distributor check any request is arrived.
3. The Distributor send data to agents.
4. The Distributor can also add additional parameters.
5. Agents is incremented by one.
6. The request may be explicit or implicit.
7. If it is implicit, a subset pi of the data set R is given.
8. If request is implicit, it is checked with the log, if any previous request is same.
9. If the request is same then the system gives the data objects that are not given to previous agent.
10. If the request is explicit ,10% tuples inserted in it are fake objects.
12. when the data gets Leaked from set E.
13. The Distributor gets an alert message when data is leaked from the agent.
14. Then Block the leak data and destroyed the data.

Algorithm for Sample Request

1: a← 0
2:R1←ϕ,……Rn←ϕ
3: remaining ← mi
4. for all i=1,……n |Ri|<mi do
5. k ←SELECTOBJECT(i,Ri)
6. Ri ←Ri^{U}{tk}
7. a[k] ← a[k]+1
8: remaining←remaining -1.

Algorithm for Explicit Request
1: R ← ∅
2: for i = 1, . . . , n do
3: if bi > 0 then
4: R ← R ∪ {i}
5: Fi ← ∅
6.while B>0 do
7: i←SELECTAGENT(R,R1,…….Rn}
7: f←CREATEFAKEOBJECT(Ri,Fi,condi)
8: Ri←R^{U}{f}
9: Fi←Fi^{U}{f}
10: bi ← bi – 1
11.if bi-0 then
12.R←R/{Ri}
13.B←B-1

Flow algorithm for data allocation strategy

Algorithm Table		
Input	Process	Output
{R1,Rr2…Rn}	Distributor sends 'n' number of data	Ri to Ai agents
{F1,F2…Fn}	Fake objects are added by unauthorized parties	Fi to Ai (Data is leaked)
{i1,i2……in}	Distributors, agents, and fake objects are incremented.	Di Ai
{k……kn}	Can select create and add the objects.	i Ri
{E……En}	Representing leaked data	Ei to 'n' number of Ai
{a……...an}	Indicates number of leaked data	Number of leaked data and agents can receive the data a=a[k]+1
{B…….Bn}	Blocking the information leaked. And then the data , reducing the data from database.	B to B-1

	Phase 1	Phase 2 Process 1	**Phase 3**	Remark
CASE1	Distributor sends data to agents →	Agent send the received data →	Authorized Parties	Case 1 is satisfied
CASE2	Distributor sends data to agents →	Agents sends data to other party →	Unauthorized Parties	Distributor logins and checks that, the agents has got the data or not and then checks the probability that which agent has leaked the data by AgentGuilt Model.

Volume : 5 | Issue : 6 | June 2015 | ISSN - 2249-555X

CASE3	From identifying the probability graph →	Then distributor tries to block leak data from unauthorized system	Case 3 is satisfied	

Data Leakage Detection comparison of papers in literature survey

Paper 1: Data Leakage detection

Step 1: In this paper technique used is "Perturbation" for detecting leakage of a set of objects or records.

Step 2: After giving a set of objects to agents, the distributor discovers some of the objects in an unauthorized place.

Step 3: At this point the distributor can access the leaked data came from one or more agents.

Step 4: Technique was used by this paper is perturbation for identifying guilty agent (leak data).

Step 5: To known how the data is leaked they have use "Probability"for identifying leak data.

Step 6: Pr(probability) Pr{Gi/S}.

Step 7: For example Probability of finding leak mails. If there are 100 mails and distributor finds only 10 mails then he used Pr{90/100} = 0.9

Step 8: Distributor find 9 mails have been leaked out of 100 mails and try to find guilty agents.

Step 9: If any agents are remaining then calculated equation is:

$$Pr\{Gi/S\} = (1-(1-p)/2)*(1-(1-p)$$

Paper 2: A Model For Data Leakage Detection

Step 1: In this section technique used is "Perturbation" distributing objects to agents, in a way that improves our chances of identifying the agents. In this algorithm.

Step 2: They have considered the options of adding "fake" objects to the distributed set. Such that objects don't correspond to the agents.

Step 3: This data make in less sensitive manner. To solve this problem algorithm used is data allocation problem setup.

Step 4: Finding guilty agents is by giving equation:-

U1 leaked t1 to S (1-p)/2

U2 leaked t1 to S (1-p)/2 with probability (1-p)/2

Step 5: They have find that U1 is guilty agents from the distributor by using equation called as Pr{Gi/S}=1-Pr{Gi/S}.

Paper 3: Data Lekage Detection and E-mail Filtering

Step 1:In this paper they have concentrated to develop a model for finding guilty agent.

Step 2: They have used "Fake Records". i.e; Which are not real but appear as a real records in order to find the guilty agent.

Step 3: To find out the guilty agent they have implemented by using fake objects acts like a watermarks like in watermarking technique.

Step 4: Which can improve over the limitations of watermarking technique as it does not required any modifications of orginal data.

Step 5: And also they have worked on e-mail filtering technique in which unauthorized users will be unable to see and download content of the e-mails which is send by the guilty agents.

Step 6: So,that distributor important data is secured.

Step 7: Guilty agent can be identified by Explicit and sample requesits. Ie; S is the set of distributor objects or data S={s1…..sn}. Set ofagents A={t1…..tn}. t1 receives objects from s1. Si is the subset of S. If the data is same from distributor and agents the agent ti is said to be an guilty agent.

Data Allocation Strategy Algorithm

tep 1: We have used technique for our paper called "Data Allocation Strategy" where in distributor sends data to agents.

Step 2: Agents sends the data to authorized parties.

Step 3: Instead of sending data from authorized parties he will send that data to unauthorized parties.

Step 4: Then distributor have to get the acknowledgement form unauthorized that data have been leaked.

Step 5: To known how the data is been leaked we have implemented first calculated using probability that is

Pr{Ri/S} For example: If there are 1000 data in database. The Distributor find the leak data only 995 so he have to divide by using probability equations.

Step 6: Suppose a distributor send a set T={t1,tm} of important data objects. The distributor send a data to agents.

Step 7: Agents gets distributor data that is sets of agents U={u1,um}

But this agent does not have wish the objects be leaked to third parties.

SStep 8: An agent Ui receives a subset € of ti (distributor) objects belongs to T, determined by a sample request or explicit request.

Step 9: Set of distributor data T={t1,t,}

Step 10: Set of agent data U={u1,um}

Step 11: Ui € Ti which belonging from T.

Step 12: These request can be sample or explicit data request.

Volume : 5 | Issue : 6 | June 2015 | ISSN - 2249-555X

Step 13: Sample request Ti=SAMPLE(T,mi): Any subset of mi records from T can be given to ui.

Step 14: Explicit request Ri=EXPLICIT(T,condi): Any ui recives all the T objects that satisfy condition.

Step 15: After giving objects to agents, the distributor discovers that a set S of T has been leaked.

Step 16: This mean some of third parties called the target has been caught in possession of S.

(S) is the leaked data in a website or other sites.

Step 17: The target termed over S to the distributor. Since the agents u1, u2 have some of the data.

Guilt Agent model

Assumption 1: For all t, t' € S such that t ≠ t' the provenance of t independent of the provenance of t.

Assumption 2: An object t € S can only be obtained by the target in one of the two ways.

A single agent ui leaked t from its own Ri set; or

The target gueesed (or obtained through other means) t without the help of any of the n agents.

In other words for all t € S the event that target guessed t and the events that agents ui{i=1……n}.

Before we present the general formula for computing Pr{Gi|S} we provide simple example

Assume that sets T,R and S are as follows:

T={t1,t2,t3}

R1={t1,t2}

R2={t1,t3}

S={t1,t2,t3}

In this case all three of the distributor's objects have been leaked and appearin S. Let us first consider how the target may have obtained object t1, which has given to both agents.from assumption the taeget either guessed t or one of ui or u2 leaked it.

We know that the probability of the former event is P, so assuming that probability that each of the two agents lleaked t, is the same we have the following cases:

Finding guilty agents is by giving equation:-

U1 leaked t1 to S (1-p)/2

U2 leaked t1 to S (1-p)/2 with probability (1-p)/2

similarly, we find that target ui, leaked t2 to S with probability 1-p since he is the only agent that has this data object.

Given these values, we can compute the probability that agent ui is not guilty agent,namely that ui did not leak either object.

Pr={Gi|S}=(1-(1-p)/2)*(1-(1-p)).

They have find that U1 is guilty agents from the distributor by using equation called as Pr{Gi/S}=1-Pr{Gi/S}.

$$\frac{\sum_{i=1}^{2} remaining}{\sum_{i=1}^{2} remaining} = re$$

Benefits

Any company will be having confidential data that should be secured so our purpose of this paper is to secure the confidential data from unauthorized party.Any confidential data of a company is very important if that important data is lost then the company value goes down,which leads to loss of data to that company.So, the information which is very beneficial should be secured very cautiously.If that data is leaked then we get an "alert message" that the data is leaked.And by this it is possible to access the possibility that an agent is responsible for the leakage.

The Information leak discovery provides data protection via;inspection and security monitoring , but in real world implementation can be more complex and expensive. Information leak discovery is the one of the most promising and least understood , security technologies to emerge during the last few years.USB, Files and Drives encryption and email filtering were as the most common information leak discovery. The Information leak discovery feature does not have its place ,particularly for clients on a budget ,or with only data protection needs. However ,the information leak discovery,includes much more effective workflow or legal will be provided. They also allow users across multiple channels,rather than defining the implementation in multiple tool. For example: when hospital may give patients records to researchers who will devise new treatments.

Conclusion

From this work, we conclude that the information leakage system model is really useful as compared to the existing system i.e watermarking technique. And we can also supply security for the data and authentication during distribution of information when it is leaked. Therefore, using data allocation algorithms and example security is not broken. Today, in this world data security is most important in various subject areas. Our model is comparatively differently and simple. The algorithm we have implemented is used in a variety of data distribution which can improve, to find the guilty agent.

REFERENCE [1] Papadimitriou and H. Garcia-Molina "Data leakage detection " IEEE Transaction on knowledge and data engineering, pages 51-63 volume 23, January 2011 | [2] P. Papadimitriou and H. Garcia-Molina, " Data leakage detection", Technical report, Stanford University, 2008. | [3] YIN Fan, WANG Yu, WANG Lina, Yu Rongwei. A Trustworthy-Based Distribution "Model for Data Leakage Detection": Wuhan University Journal Of Natural Sciences. | [4] P. Buneman , S. Khanna and W.C. Tan. Why and where:" A characterization of data provenance". ICDT 2001, 8th International Conference, London, UK, January4-6, 2001, Proceedings, volume 1973 of Lecture Notes in Computer Science, Springer, 2001. | [5] S. Czerwinski, R. Fromm, and T. Hodes. "Digital music distribution and audio watermarking". |