A Survey on Data leakage Optimization and Prevention by Identifying Guilty Agent without Causing Disturbance and Inconvenience to Trusted Agent

Sangramsinh B. Deshmukh[1], R. N. Phursule[2]

[1]Department of Computer Engineering, JSPM's Imperial College of Engineering & Research, Wagholi, Pune, Maharashtra, India

[2]Professor, Department of Computer Engineering, JSPM's Imperial College of Engineering & Research,
Wagholi, Pune, Maharashtra, India

Abstract: ***Data misuse might be performed by peoples or sources like an organization's employees or business partners who have access to sensitive data and misuse their authority. We can say that users are either trusted or untrusted. The access privilege to untrusted parties to data objects (e.g., patient records or clients) should monitor to detect misuse. Still, monitoring data collection or data information is resource dependent and it taking much more time and may also cause issues like disturbance and inconvenience to involved employees. So that, monitored data collection or information should selected carefully. In this paper, we represent two optimization issues designed carefully for fetching specific data collection or information for monitoring, such that the detection rate is maximized and the monitoring effort is reduced. In the first , the goal is to select data objects for monitoring that are accessed by at most c trusted agents while ensuring access to at least k monitored objects by each untrusted agent (both c and k are integer variable). The goal of the second is to select monitored data objects that maximize the number of monitored data objects accessed by untrusted agents while ensuring that each trusted agent does not access more than d monitored data objects (d is an integer variable).***

Keywords: Security, Data misuse, data monitoring, honeytokens, information security, Guilty agent, trusted agents

1. Introduction

Data leakage is the most important security threat to organizations. Data leak became very dangerous issue, particularly when data leak is done by law representatives like agents. In this paper, we shown the possible techniques for finding and valuing the wrong doing of representatives and we will differentiate trusted and untrusted agents. Water marking is the most used technique for data leakage detection and guilty agent findings which may cause some modifications to the data. To overcome the issue of using watermark, data allocation strategies are used for detecting guilty agents. Distributor smartly and intelligently allocates data objects depending on sample request and explicit request by using allocation strategies in order to better in sensing responsible person or guilty agent. Fake data objects are designed to look like true data objects, and are supplied to agents together with requested data objects. Fake data objects are encrypted with a private key and are designed to look like true data objects, and are distributed to all agents together with requested data. By this way we can find out, the responsible person i.e. a guilty agent who leaked the private data by decrypting his fake object.

Data leakage can take place through a range of methods - some are easy and simple where as some are complex. So, there is no single and only technique to control leakage of data. Now days, Data leakage detection has became most important part of any organizations capability and ability to manage and keep safe these important and confidential information. examples of critical and to be kept secret information that applications may contains : to do with IPR(Intellectual Property Rights), Corporate private information and data and customer or stakeholder information and data .Watermarks are very useful in knowledge-base , which includes some adjustment and modification of data. Purpose of our Paper is to discover when the distributor's sensitive data or information has been leaked by representatives' i.e. guilty agents, and probably identify the guilty agent that leaked the data using encrypted fake data objects and while doing this all trusted agents should not get affected .

2. Literature review

Our scenario is based on the scenario presented by Papadimitriou and Garcia-Molina[2010], with several modifications. In their paper, Papadimitriou and Garcia presented a method for data leakage detection. In the scenario that they address, a distributor distributes sensitive data to almost all agents according to a specific request that is issued for each one of the agents. An example of such a scenario is a proactive CRM system in which the data owner decides which customer or stakeholder to call, and the customer or stakeholder details are forwarded to the third-party call agent. If sensitive data is leaked, the data owner would like to be able to identify the source of leakage, or at least be able to estimate the likelihood of each agent to have been involved in the incident. Therefore, A guilt model is proposed for estimating the probability that an agent is involved in a given data leakage. The capability and ability to identify the of the leakage depends on the distribution of data objects among the agents. Therefore, a data allocation method that distributes data records among the agents based on the agents' requests and optimization models are presented. The proposed allocation method ensures that object sharing among the agents is minimal, and therefore, in the case of a leakage incident, the data owner will be able to use the guilt model to identify the source of leakage with high probability..

Two types of data requests are being considered in Papadimitriou and Garcia-Molina [2010]: explicit request and sample request. An explicit request contains predefined conditions, and all of the objects in the dataset that comply with these conditions must be returned. A sample request defines the amount of objects to be randomly selected from the entire dataset. Combined requests (i.e., requests for a sample of objects that comply with a predefined condition) are not handled by the proposed algorithms; However, it is explained how they might be handled using the proposed algorithms. They as well proposed including fake data to the lists of real data objects when distributing them to the agents. Fake data objects may help to better distinguish between the agents and increase the accuracy of the guilt model (e.g., when each untrusted agent receives a unique fake object).

Four scenarios can be stated by the two request types (sample or explicit) and the two options of planting fake objects in the result sets (using or not using fake objects). It is assumed that in each scenario, all of the agents issue the same type of requests (i.e., either explicit or sample queries), and if fake objects are up to use, the same amount of objects will be planted for all agents. Several allocation algorithms are proposed to deal with each scenario. Empirical evaluation showed that the proposed algorithms reached a significantly greater ability to identify the source of leakage (compared with simple allocation algorithms), even in cases where there was a large overlap between the objects that the agents received.

3. Proposed Method

We extend the scenario presented by Papadimitriou and Garcia-Molina [2010] (i.e., A data owner distributes objects among agents based on predefined queries) with the following four main differences.

1. *Trusted and Untrusted Agents*. While all of the agents in Papadimitriou and Garcia-Molina [2010] are considered to be untrusted, the agents in our scenario are divided into two types: trusted agents and untrusted agents. For a specific group of agents, trusted agents are usually the group's members, and untrusted agents are those who do not belong to the group. The designation as trusted or untrusted agent depends on the relationship to the data objects accessed. For example, internal employees of the organization may be considered as trusted while delegates of third party and business partners will be considered as untrusted, or the employees of one department (e.g., marketing) may be considered as untrusted when referring to the data that is under the responsibility of another department in the same organization (e.g., Information technology).
2. *Detecting Data Misuse Events*. The objective of Papadimitriou and Garcia-Molina [2010] is to identify, with a high probability, the source of a data leakage. This is done by finding an optimal allocation of the data objects among the agents and not by actively monitoring the interaction of agents with the data. We consider that to detect a data misuse action, the organization needs to provide in-depth monitoring and investigation of the activity of an agent on a data object. Since monitoring all actions performed over all data objects is a complicated and expensive task, identifying the most important and beneficial data object for monitoring is needed. So, we would like to identify the optimal set of data objects that should be monitored to detect data misuse actions. The objects selected for monitoring will be those accessed for the most part by untrusted agents, with access by few trusted agents. Our goal is to maximize the detection rate of data misuse events performed by untrusted agents while minimizing the monitoring effort.
3. *Given Allocation of Data Objects*. In the case that we are exploring, a list of accessible data objects (from the entire set of objects) is defined for each agent (both trusted and untrusted). However, since we are not dealing with allocation problems but rather are selecting objects for monitoring after the allocation has been done, the request type of each agent (explicit or sample) does not influence our solution and therefore is not relevant.
4. *Planting Honeytokens*. Similar to Papadimitriou and Garcia-Molina [2010], we consider the option of planting fake objects, also known today as honeytokens, in the object list provided to each agent. In our scenario, honeytokens can be selected in a way that each untrusted agent's list will contain at least a predefined number of honeytokens while minimizing the chance that these honeytokens will be included in the trusted agent's lists as well.

To demonstrate the new scenario, we will use the following example of a small bank in which some of the internal employees (considered trusted agents) are serving as clerks with different fields of expertise. Each trusted clerk has a group of customers for which he is allowed to access data, based on his relevant field of expertise. Moreover, the bank shares information about its customers with several credit card companies to provide joint services to its customers. The credit card companies' delegates are considered untrusted agents by the bank. In our example, each customer account can be referred to as a data object.

4. Conclusion

This paper studies the problem of efficiently detecting data leakage and guilty agents in very large observation databases collected by systems. We present a method for selecting specific data objects to efficiently monitor and detect data misuse incidents performed by insiders. In the addressed scenario, trusted and untrusted agents are authorized to access a predefined list of data objects out of a shared data object collection. Our method suggests monitoring only a subset of data objects that are selected in such a way that the monitoring effort is minimized while the detection rate is maximized.

References

[1] Panagiotis Papadimitriou, Hector Garcia-Molina , IEEE Paper “Data Leakage Detection”,2011.

[2] Panagiotis Papadimitriou, Hector Garcia-Molina , IEEE Paper “Data Leakage Detection”,2010.

[3] Asaf Shabtai, Maya Bercovitch, Lior Rokach, and Yuval Elovici. 2014. Optimizing data misuse detection. ACM Trans. Knowl. May 2014.

[4] P. Buneman, S. Khanna, and W.C. Tan, “Why and Where: A Characterization of Data Provenance,” Proc. Eighth Int’l Conf. Database Theory (ICDT ’01), J.V. den Bussche and V. Vianu, eds., pp. 316-330, Jan. 2001.