A Review on Tracing Data Provenance in Malicious Environments

Neha Belekar, R. P. Dahake

[1,2]*Dept. of Computer Engineering, MET's Institute Of Engineering Nasik, Maharashtra, India*

---***---

Abstract - *In today's era, information leakage is one of the most serious threats to companies. A data owner sends secret or confidential information to a group of trusted agents. Some of the information is lost and found in an inappropriate place. Thus data is leaked. Data leakage means data distributed by the data owner is leaked by one or more agents. This causes a huge harm to the business. The distributor must assess whether data is leaked from one or more agents. To improve the probability of identifying leakages data allocation strategies (across the agents) are used. A data lineage framework is used for identifying a guilty entity. The digital watermarking is a technique in which vital information is kept hidden in the original data for protecting unauthorized duplication and distribution of data. An accountable data transfer protocol is built using oblivious transfer, robust watermarking, and signature primitives. In some occasions fake data records are injected in order to improve detecting data loss and identifying the guilty entity. The data sent by the data owner must be protected, secret and it must not be regenerated. The framework of data lineage is considered for transmission of data and is a key step towards achieving accountability.*

***Key Words*: Information leakage, data provenance, accountability, watermarking, distributor, agent**

1. INTRODUCTION

In the course of technology and doing business, at many occasions sensitive important data is handed over to trusted third parties. For example, an organization may have associations with other organizations that share customer data. Another business enterprise may outsource its work to other companies where they require to send data to an external company. The owner of the data is called as the data owner or distributor and the third parties are called as the agents or data consumer. The aim is to detect leakage of sensitive data and identification of the guilty agent. In a very short time, large volumes of digital sensitive data can be copied by the attacker and can be spread across the internet. Currently there is no accountability method, the risk of getting caught for data provenance is very low. Nowadays, due to these reasons, the problem of data leakage has reached a new height.

It is a concern for not only organizations, but individuals are also affected by data provenance. The situation get worse because of the advent in the technology of mobile phones and social networking. In today's environment, individuals usually expose their personal data to various service providers, in return for some services which are free of cost. In the nonappearance of right regulations and accountability methods, many of these online applications share entities identifying information with couple of advertising and online tracking companies. Even with access control methods, where there is restricted access to confidential data, a malicious attacker can publish confidential data as soon as he receives. Information security mechanisms like encryption offer protection as long as the information is encrypted, but once the consumer decrypts a message, nobody can prevent the disclosure of decrypted content. Thus it seems impractical to avoid data leakage fanatically.

Data provenance is the enormous threat in front of the companies and various different enterprises. Though there are number of different encryption mechanisms designed for securing information, there is a challenging problem of the integrity of the users of the systems. In order to offer security against data loss threat technologies like machine learning content/context based detectors, encryption, access control, firewalls and identity management have already been incorporated.

The information distributed is considered as sensitive data when it consists of information about the client, budget, code and any design specification. The agents who get their hands on the sensitive data are also known as cyber criminals. Data leakage is done for their own profits which results in loss of the company. To overcome this problem, a general method of data transmission is used. This mechanism is referred as accountability.

This accountability method can be directly correlated with detection of data transfer history across multiple nodes right from its origin. The distributor sends the data to the agent using strategies that increase the possibility of finding the agent by adding fake data to the information distributed. If any person receiving the data leaks the data then the distributor will find the agent by the help of number of fake objects released out and the distributor waits until he gets enough evidence and finally conform the agent and closes the business with him or takes any legal action on the agent.

2. LITERATURE SURVEY

G. Doerr et al. [1] noted that digital watermarking recently elongated from still images to video content. Further research in this area is strongly encouraged by an increasing requirement from the copyright owners in order to protect their rights assuredly. A watermark can be divided into two parts: one for copyright protection and the other for customer fingerprinting. Robustness has to be considered attentively. Applying watermarking to video is definitely a new area of research by making use of still images. The easier and simpler approach is to consider a video as a series or chain of still images. An existing watermarking method can be reused for still images. A new robust video watermarking algorithm can be designed by exploiting additional temporal dimension. Another approach considers a specific video compression standard which can be used to compress a video stream.

M. A. Alsalami et al. [2] proposed the technique of digital audio watermarking for embedding data along with audio signal. Copyright owner uses embedded data for identification purpose. The main aim of watermarking systems is to embed a hidden robust watermark into digital media file. These systems have to appease two contrary needs. First, watermark must be unaffected from voluntary and involuntary removal. Second, watermarked signal should preserve a fair loyalty and robustness of watermarked signal.

V. M. Potdar et al. [3] noted, a lot of research is being conducted currently in the field of watermarking. There is a lot of work begin conducted indifferent branches in this field. Steganography is changing the image in a way that only the sender and the intended recipient is able to detect the message sent through it. Steganography is used in modern printers. It has been used allegedly by terrorists and intelligence services Watermarking is used to verify the identity and authenticity of the owner of a digital image. It is a process in which the information which verifies the owner is embedded into the digital image or signal. These signals could be either videos or pictures or audios. For example, famous artists watermark their pictures and images. If somebody tries to copy the image, the watermark is copied along with the image. It is used for copyright protection, source tracing, and annotation of photographs.

A. Mascher-Kampfer et al. [4] discussed about the purpose of single watermarking schemes in a multiple re-watermarking scenarios. A surprisingly huge number of different watermarks may be detected and also robustness can be preserved up to a certain limit using this approach, however, detection association falls for an increasing number of embedded marks which restricts scalability to long selling series.

R. Halder et al. [5] improved and proposed digital watermarking for relational databases. It arrived as an answer in providing protection, intrusion detection, culpable tracing, and maintaining integrity of relational data. The current modern techniques for relational databases are fingerprinting and watermarking. All the techniques classified are based on (i) whether the technique introduces the distortion to underlying data, (ii) the type of the cover where mark is embedded, and (iii) the type of the watermark information. For safeguarding the ownership distortion-based watermarking methods are practiced. And for maintaining integrity of the database distortion-free watermarking methods are adopted.

P. Papadimitriou et al. [6] initiated the study of data leakage in which they defined that data owner has given confidential data to a group of data consumers (third parties). In some cases the data is lost and found in an inappropriate place (e.g., on the internet or somebody's mobile or computer). In order to improve the probability of detecting data leakage certain approach of allocating data (across the agents) is used. In some situations addition of fake data records will further enhance the likelihood of detecting leakage and identifying the culpable entity.

Michael Backes et al. [7] in this work, a LIME framework is implemented having two roles data distributor and agent. The data lineage approach is designed to identify accusable entity, and identify the non-repudiation and honesty assumptions. A data transfer protocol is used between distributor and agent within a malicious environment. Oblivious transfer, robust watermarking, and signature primitives' mechanisms are developed and analyzed.

3. SYSTEM FRAMEWORK

A.Generic Data Lineage Framework

The generic data provenance framework is incorporated for flow of data across multiple nodes or entities in the malignant surroundings. Identification of an optional non-repudiation assumption between two distributors, and an optional trust (honesty) assumption is done by the auditor about the distributors. Non-repudiation assumption: If the files are transmitted from one distributor to another distributor, then it can be pretended that the transmission is administered by a non-repudiation assumption. Sending owner trusts the receiving owner to take the responsibility if he leaks the document. Fingerprinting: It is a method to embed identifiers into files for uniquely identifying the recipient. Honesty Assumptions: Data distributor does not expose the file and accuses another agent for the leaker.

B. Accountable Data Transfer Protocol

The protocol specifies how one data owner sends a file to another one, what message or data is embedded and which steps the data auditor executed to search the culpable entity in occasion of data exposure. In such circumstances, it is assumed that both entities involved should know each other's signature verification key. The major functionality of this model is that it prosecutes accountability by design.

C. Roles involved in Lineage Framework

The three important parts or roles that can be assigned to the involved parties in lineage framework: data owner (distributor), data consumer (agent) and auditor (controller).
Data Owner: The distributor manages the files, documents and the agent is the recipient of the documents and can achieve some task or work by making use of the files or documents received.
Data Consumer: which receives the document. An agent can send a file to another agent, so we also have to examine the case of an untrusted sender. Every agent can expose new embedded information to the controller to indicate the next agent and to justify his own guiltlessness.
Auditor: which is not involved in the sending of documents or files, it is only call forth or requested when an exposure occurs and then executes all steps that are necessary to identify the accused entity.

Entire set of quoted roles can have many occurrences. In classic scenarios the distributor sends files to agents. However, there is a probability that agent can circulate the files or documents to other consumers or that distributors swap documents with each other. In the outsourcing or redistribution plot the employees and their employer are distributors, while the outsourcing organizations are doubtful agents. Below given Fig 1 shows the framework for tracing data lineage.

D. Processing Steps

The procedure to find the accusable entity proceeds in the following manner:
1. Firstly the data owner will select an input file and the agent to whom the file will be sent.
2. Then, the auditor will embed the watermark by collecting the original file and name of the recipient from the distributor.
3. Fake objects are added and the watermark signature is placed into the file by applying accountable transfer protocols.
4. Further, in order to decode the file, it will be collected, loaded, analyzed and decoded.
5. Guilty attacker is founded by tracing the data lineage and leakage.
6. Finally, auditor will detect and identify the guilty party involved in the data transmission.

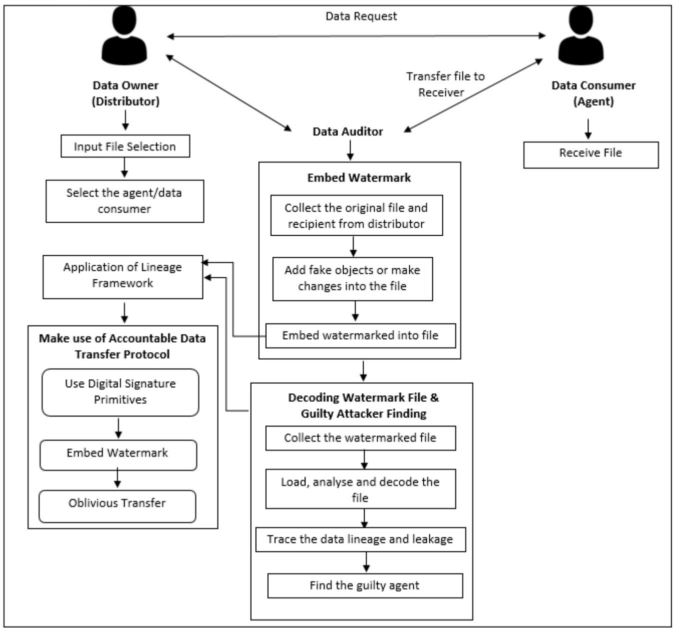
Data Request
Data Owner (Distributor)
Transfer file to Receiver
Data Consumer (Agent)
Data Auditor
Input File Selection
Receive File
Embed Watermark
Select the agent/data consumer
Collect the original file and recipient from distributor
Application of Lineage Framework
Add fake objects or make changes into the file
Make use of Accountable Data Transfer Protocol
Embed watermarked into file
Use Digital Signature Primitives
Decoding Watermark File & Guilty Attacker Finding
Collect the watermarked file
Embed Watermark
Load, analyse and decode the file
Oblivious Transfer
Trace the data lineage and leakage
Find the guilty agent

Fig 1: Framework for tracing data provenance

4. CONCLUSIONS

The chances that a data consumer is culpable for data loss is checked on the basis of overlay of his data with the exposed data and the data of other consumers, and based on the possibility that data items can be presumed by other modes. The lineage approach appliances a wide range of data circulation methodologies that can boost the owner's likelihood of finding leakage and diagnosing a data leaker. Thus the data provenance model is effective than the existing watermarking model. Data provenance model caters protection to data at the time of circulation or transmission of data and can also find if that gets leaked. Watermarking safeguards the data using techniques like encryption, whereas data provenance model provides prevention plus guilt identification. This model proves to be advantageous to enterprise, where data is disbursed using any public or private medium and shared with the outsider (third party). Now, enterprise, numerous organizations can entrust or depend on this data provenance model.

ACKNOWLEDGEMENT

We are glad to express our sentiments of gratitude to all who rendered their valuable guidance and for their extended help & suggestions at every stage of this paper. We would like to express our appreciation and thankful to Dr. M. U. Kharat, Head of Department, Computer Engineering., MET's IOE, BKC, Nasik. We also thank the anonymous reviewers for their comments.

REFERENCES

[1] G. Doerr and J.-L. Dugelay, A guide tour of video watermarking, Signal Process.: Image Commun., vol. 18, no. 4, pp. 263 282, 2003.

[2] M. A. Alsalami and M. M. Al-Akaidi, Digital audio watermarking: Survey, School Eng. Technol., De Montfort Univ., U.K, 2003.

[3] V. M. Potdar, S. Han, and E. Chang, A survey of digital image watermarking techniques, in Proc. 3rd IEEE Int. Conf. Ind. Informat., pp. 709716, 2005.

[4] A. Mascher-Kampfer, H. Stogner, and A. Uhl, Multiple re-watermarking scenarios, in Proc. 13th Int. Conf. Syst., Signals, Image Process., pp. 5356,2006.

[5] R. Halder, S. Pal, and A. Cortesi, Watermarking techniques for relational databases: Survey, classication and comparison, J. Universal Comput. Sci., vol. 16, no. 21, pp. 31643190, 2010.

[6] P. Papadimitriou and H. Garcia-Molina, Data leakage detection, IEEE Trans. Knowl. Data Eng., vol. 23, no. 1, pp. 5163, Jan. 2011.

[7] Michael Backes, Niklas Grimm, and Aniket Kate Data Lineage in Malicious Environments, in IEEE Trans.Dependable and Secure Computing,vol. 13, no. 2,Apr. 2016.