ISSN: 2321-7782 (Online)
Volume 4, Issue 1, January 2016
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Detection of Data Lineage with 1-out-of-n way

Sonali Vijay Patil[1]
Department of Computer Engineering,
Savitribai Phule Pune University,
Pune, India

Baban Thombre[2]
Department of Computer Engineering,
Savitribai Phule Pune University,
Pune, India

Abstract: ***The authorize or unauthorized leakage of secret data is no doubt one of the most major security problems which organizations or systems face in this era. It also affects our personal day to day life: The personal information is available on social networks, or now-a-days it is also available on Smartphone is intentionally or unintentionally transferred to third party or hackers. Also a data distributor may give confidential data to some trusted agents or third parties. During this process some data is leaked or transferred to unauthorized place. We propose data allocation strategies that will give more probability of identifying leakages. We present a LIME data lineage framework for data flow across various locations.***

By using oblivious transfer, robust watermarking, and signature primitives we develop and analyze the data transfer protocol in a malicious environment between two entities. At the end of we perform an experimental result and analysis of our framework.

Key words: Watermarking,Data leakage, oblivious transfer.

I. INTRODUCTION

The authorize or unauthorized leakage of secret data is no doubt one of the most major security problems which organizations or systems face in this era. It also affects our personal day to day life. The Privacy Right Clearinghouse in the United States maintains the chronology of data braches. They found that from 4355 data braches near about 868,045,823 records are brached which made public since 2005. So the loss of data or leak of data will cause loss to the organization or most of the companies. Also the organization have fear of losing customers confidence, their support or maybe they have to pay fine for data loss. For all these reasons, data leakage becomes headache to organization.

Also the management of data is become crucial. In the metadata we can again and again use the same data for our purpose. Now we are using data provenance, which is like metadata concern to data product from original locations. Data leakage is defined as an unofficial, unauthorized transfer of data, information from computer to the outside world and data lineage is defined as chain of data which includes origin of data and where it goes over time. Data lineages do the survey of how this information is used and also track the data system. Data lineage gives the data source and intermediate data flow hops with backward data lineage which finally gives intermediate data flow hops with forward data lineage.

The loss of data is not only concern the organizations but also concern to each person. In these scenarios, the person share or disclose their personal information to outside world, third party because of some part of profit to them. They will transform the information through various social networking sites like face book, twitter, LinkedIn and so many and through smart phones also. Afterwards this data through illegal manner may share the individual's personal information to lots of marketing, advertising, and internet tracking companies.

For example, a hospital may give patient details to researchers who may find the new diagnosis pattern. Enterprise may transfer or outsource data to other companies. So the main objective is to detect when the data is transferred from owner to receiver have been leaked by agents, and also identify the agent who investigate the data.

II. Overview of Data Lineage

Data lineage describes or gives presentation on the data process with various changes from sender to receiver or from source to destination in the enterprise environment. Data lineage will give brief idea about how the data flows along the route, how the changes are made, it will show about what changes are made at each stage, how the data separates after each stage, gives simpler and easier outline of data lineage by dots and lines. Dot means a data container for data points and lines connecting them shows data point undergoes between data containers.

Need data lineage:

Data leakage may takes place in most of the field ,therefore it is necessary to identify data leakage types. Data leakage takes place by following users:

1. The illegal users
    - Company IT resources used in manner that they shouldn't, it means by saving movies, songs, and playing games
2. Illiterate
    - Employees with some or lack of knowledge of security
    - Risk due to violation of data
3. The fed up employees
    - Shortage of employees
    - They shouldn't send official data or information to third parties by gaining access to IT system.
4. Appliances
    - Inserting many kind of applications to their work PC's
    - Download Software

Data Leakage Prevention:

- Build security mechanisms like Firewalls, IDS, antivirus software and thin client architecture.
- Use of reasoning algorithms and pattern based monitoring tools.
- Use of access and device control.
- Encryption keys storage.

Challenges in Data Lineage:

- Lineage store scalability
- Fault Tolerance
- Capture of Lineage for Black Box operators
- Efficient Tracing
- Sophisticated Replay
- Anomaly Detection

III. LITERATURE SURVEY

- Panagiotis Papadimitriou and Hector Garcia- Molina of Stanford University they develop model for data leakage detection for accessing guiltiness of agents. The data distributor to identify the malicious agent which leaked the information. In addition, they argue that current watermarking techniques are not practical, as they may embed extra information which could affect agents' work and their level of robustness may be inadequate.

  Cons: Existing matchmaker algorithm is unable to take correct decision based on QoS parameters.

- Hasan, Sion and Winslett- They summarizes the data provenance. They stated that a system that enforces logging of read and write actions in a tamper-proof provenance chain. This creates the possibility of verifying the origin of information in a document.

  Cons: However, as an attacker is able to strip of the provenance information of a file, the problem of data leakage in malicious environments is not tackled by their approach.

- A. Pretschner, M. Hilty, F. Sch¨utz, C. Schaefer, and T. Walter employ data usage control enforcement systems and preventive measures to ensure that data is transferred in distributed systems in a controlled manner preserving the well defined policies.

  Cons: Handling global constraints, can lead to poor performance rendering inappropriate for applications with dynamic and real time requirements.

  N. P. Sheppard, R. Safavi-Naini, and P. Ogunbona find out the problem of an insider attack, where the data generator consists of multiple single entities and one of these publishes a version of the document. Usually methods for proof-of-ownership or fingerprinting are only applied after completion of the generating process, so all entities involved in the generation process have access to the original document and could possibly publish it without giving credit to the other authors, or also leak the document without being tracked.

- Cons: The problem can be solved by the usage of watermarking and possibly even by using complete fingerprinting protocols during the generating phase of the document.

Proposed System:

By Using Forensic techniques we can identify leakage in malicious environment but it is cost effective. That is it is beyond the desirable cost and also doesn't generate the appropriate results. We have to solve the leakage of data problem which occur in many scenarios by postulating to find out the provably associating the guilty party to the leakages and work on the lineage methodologies. For that we define LIME, which is lineage in malicious environment, which shows the flow of data through various locations.

Lime Framework:

For any kind of data we can use LIME (Lineage in Malicious Environment) Framework by using Watermarking scheme. For that we have to know the watermarking techniques for different data types. Lime should be used in all cases as it is general model. We extract the data type and every data item called as document. It involved characteristics or agents as: data owner, data consumer and data auditor. Data owner is an entity which takes part in management of documents and it is responsible for its accuracy, integrity and timeliness, also it can authorize or deny access to use of data at certain limit. Data consumer will receive the data or documents from data owner and carry out some tasks using them, also the data resources are consumed by them by using data set. Data auditor is act as a third party which conducts the processing of data audit and access the data to fit for given purpose. He is not involved in sharing of documents and only invoked when data leakage occurs.

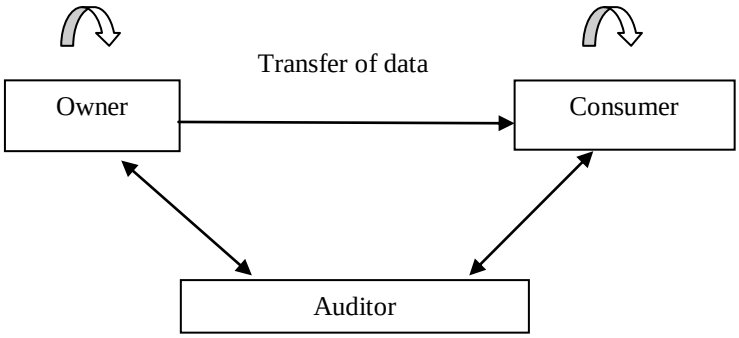
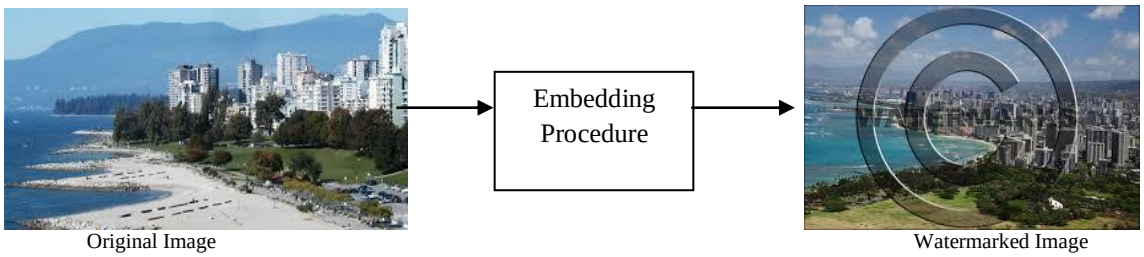


Fig1.Simple LIME Framework

Use of Watermarking:

Watermarking schemes are used in multimedia files such as audio files, images and video. They identify the image or pattern of paper which occurs in various shades as dark or light when transmitted by transmitted light. They use the wavelet or fourier transform, discrete cosine representation by using multimedia files by embedding watermarking. We can apply LIME to user database or medical records as watermarking also supports data types like relational database, android applications and text files. We create encrypted watermark to protect our data from leakage. Now for our purpose we use watermarking of text by inserting or embedding information in changing the texts appearance, we can change the distance between words and lines or insert text into the image so that it can be invisible or inaudible to humans. Also we can use language watermarking scheme which does not work on appearance, it works on semantic level of data.



Original Image

Watermarked Image

Fig. 2 Use of Watermarking

Watermarking techniques does not look in syntax they are encoded in execution state of the application which make the system robust. Therefore it will avoid misuse of data from attacker also no any other wrong entries are added in framework. The above techniques we can use in our framework by distributing splitting algorithm.

IV. APPLICATION

- It involves study of unobtrusive techniques for detecting leakage of a set of objects or records.
- 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 the proposed approach, a model is developed for assessing the guilt of agents.
- The algorithms are also presented for distributing objects to agents, in a way that improves the chances of identifying a leaker.
- Finally, the option of adding fake objects to the distributed set is also considered.

V. Conclusion

We represent LIME Framework by using watermarking techniques for sharing of data from sender to receiver across multiple locations. We can use combination of data transfer protocol, oblivious transfer and digital signature for data transfer in an encrypted Framework. LIME will determine the malicious programs who leaked the personal information or documents and provide the appropriate action to protect our data.

We prove its correctness and show that it is realizable by giving microbenchmarking results. By presenting a general applicable framework, we introduce accountability as early as in the design phase of a data transfer infrastructure.

VI. Future Scope

Our work give idea about to use different kind of data leakage techniques for various scenario and data types in future research. It will be an interesting future research direction to design a verifiable lineage protocol for derived data.

Acknowledgement

I am profoundly grateful to Baban Thombre for his guidance and continuous encouragement throughout to see that this paper rights its target since its commencement to its completion. I would like to express deepest appreciation towards his invaluable guidance and support me in completing this paper. Also I must express my sincere gratitude to all the staff members of Computer Department who helped me directly or indirectly during this course of work.

References

1. Offshore outsourcing, http://www.computerworld.com/s/article/109938/Offshore outsourcing cited in Florida data leak.
2. A. Mascher-Kampfer, H. St ¨ogner, and A. Uhl, “Multiplere-watermarking scenarios, in Proceedings of the 13th International Conference on Systems, Signals, and Image Processing (IWSSIP 2006).Citeseer, 2006, pp. 53–56.
3. P. Papadimitriou and H. Garcia-Molina, “Data leakage detection,Knowledge and Data Engineering, IEEE Transactions on, vol. 23, no. 1, pp. 51–63, 2011.
4. Pairing-Based Cryptography Library (PBC), http://crypto.stanford.edu/pbc.
5. I. J. Cox, J. Kilian, F. T. Leighton, and T. Shamoon, “Secure spread spectrum watermarking for multimedia, Image Processing, IEEE Transactions on, vol. 6, no.12, pp. 1673–1687, 1997.
6. Bhamare Ghanashyam,Desai Kiran,Khatal Supriya, Mane Vinod,Prof. Hirave K.S.,” A Survey Paper on Data Lineage in Malicious Environments” Multidisciplinary Journal of Research in Engineering and Technology, Volume 2, Issue 4,Pg.720-724
7. “Chronology of data breaches,”http://www.privacyrights.org/data-breach.
8. “Data breach cost,” http://www.symantec.com/about/news/release/article.jsp?prid=20110308 01.
9. “Privacy rights clearinghouse,” http://www.privacyrights.org.
10. Michael Backes, Niklas Grimm, and Aniket Kate,” Data Lineage in Malicious Environments” DOI 10.1109/TDSC.2015.2399296, IEEE Transactions on Dependable and Secure Computing
11. M. Handley and J. Crowcroft. Network Text Editor (NTE): A scalable shared text editor for the MBone. In ACM SIGCOMM, pages 197–208, Cannes, France, 1997.
12. F.Hartung and B. Girod. Fast public-key watermarking of compressed video. In IEEE International Conference on Image Processing, pages 528– 531, Santa Barbara, USA, 1997.
13. T. Kalker, G. Depovere, J. Haitsma, and M. Maes. A video watermarking system for broadcast monitoring In IS&T/SPIE Conference on Security and Watermarking of Multimedia Contents, pages 103–122, San Jose, USA, 1999.
14. B. Koh and T. Chen. Progressive browsing of 3D models. In IEEE Workshop on Multimedia Signal Processing, pages 71–76, Copenhagen, Denmark, 1999.
15. G. C. Langelaar, I. Setyawan, and R. L. Lagendijk. Watermarking digital image and video data: A stateof-the-art overview. IEEE Signal Processing Magazine, pages 20–46, September 2000.

