ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Data Tracing For Protecting Data Leakages in Malicious Environments

V.S.Viswanath
M.Tech CNIS Student,
Sreenidhi Institute of Science and Technology.

K.Sri Laxmi
Assistant Professor, Dept of IT,
Sreenidhi Institute of Science and Technology.

Abstract:
Data lineage is a mechanism that tells about data life cycle that includes the data's sources and where it moves from time to time. It explains what happens to data as it goes through distinct processes. It helps to contribute visibility into the analytics pipeline and simplifies tracking errors back to their resources. It also replays specific portions of inputs of the dataflow for step-by-step debugging or reestablishing lost output. In fact, database systems have used such data, called data provenance, to acknowledge similar validation and debugging objections already. Data Leakage is the illegitimate transmission of data (or information) from origin to an external target.

In this paper, we examinea generic data lineage framework LIME for data flow over multiple entities that take two characteristic, principal actors (i.e., owner and consumer). We specify the exact security guarantees prescribed by such a data lineage system regarding identification of a guilty person, and the simplifying non-repudiation and honesty assumptions.

We then develop and examine an innovative responsible data transfer protocol between two entities within a malicious environment by constructing over unaware transfer, robust Watermarking, and signature primiti. Lastly, we execute an investigational assessment to validate the practicality of our protocol and apply our framework to the important data leakage situations of data outsourcing and social networks.

Keywords:
Data Lineage, Data Leakage, Data Transfer, Watermarking, Signatures, Data Flow.

Introduction:
Data Lineage gives a visual representation to determine the information flow/movement from its origin to destination via different changes and lops on its way in the enterprise background. Data lineage represents: how the data lops among different data points, how the data gets transformed along the way, how the representation and parameters change, and how the data splits or converges after each hop. Easier representation of the Data Lineage can be shown with dots and lines, where dot represents a data container for data point(s) and lines connecting them represents the transformation(s) the data point under goes, between the data containers. Representation of Data Lineage broadly depends on scope of the Metadata Management and reference point of interest.

Data Lineage provides sources of the data and intermediate data flow hops from the reference point with backward data lineage, leads to the final destination's data points and its intermediate data flows with Forward data lineage. These views can be combined with End to End Lineage for a reference point that provides complete audit trail of that data point of interest from source(s) to its final destination(s). As the data points or hops increases, the complexity of such representation becomes incomprehensible.

Thus, The best feature of the data lineage view would be to able to simplify the view by temporarily Masking unwanted peripheral data points. A tool that have the masking feature enables scalability of the view and enhances analysis with best user experience for both Technical and business users alike.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Scope of the data lineage determines the volume of metadata required to represent its data lineage. Usually, Data Governance, and Data Management determines the scope of the data lineage based on their regulations, enterprise data management strategy, data impact, reporting attributes, and critical data elements of the organization. Data Lineage provides the audit trail of the data points at the lowest granular level, but presentation of the lineage may be done at various zoom levels to simplify the vast information, similar to the analytic web maps. Data Lineage can be visualized at various levels based on the granularity of the view. At a very high level data lineage provides what systems the data interacts before it reaches destination.

As the granularity increases it goes up to the data point level where it can provide the details of the data point and its historical behavior, attribute properties, and trends and Data Quality of the data passed through that specific data point in the data lineage. Data Governance plays a key role in metadata management for guidelines, strategies, policies, implementation. Data Quality, and Master Data Management helps in enriching the data lineage with more business value. Even though the final representation of Data lineage is provided in one interface but the way the metadata is harvested and exposed to the data lineage User Interface (UI) could be entirely different. Thus, Data lineage can be broadly divided into three categories based on the way metadata is harvested:Data lineage involving software packages for structured data, Programming Languages, and Big Data.

Data lineage expects to view at least the technical metadata involving the data points and its various transformations. Along with technical data, Data Lineage may enrich the metadata with their corresponding Data Quality results,Reference Data values, Data Models, Business Vocabulary, People, Programs, and Systems linked to the data points and transformations. Masking feature in the data lineage visualization allows the tools to incorporate all the enrichments that matter for the specific use case.

Metadata normalization may be done in data lineage to represent disparate systems into one common view.

Related Work:

Secure Spread Spectrum Watermarking for Multimedia- AUTHORS: I. J. Cox, J. Kilian, F. T. Leighton, and T. Shamoon

This paper presents a secure (tamper-resistant) algorithm for watermarking images, and a methodology for digital watermarking that may be generalized to audio, video, and multimedia data. We advocate that a watermark should be constructed as an independent and identically distributed (i.i.d.) Gaussian random vector that is imperceptibly inserted in a spread-spectrum-like fashion into the perceptually most significant spectral components of the data. We argue that insertion of a watermark under this regime makes the watermark robust to signal processing operations (such as lossy compression, filtering, digital-analog and analog-digital conversion, requantization, etc.), and common geometric transformations (such as cropping, scaling, translation, and rotation) provided that the original image is available and that it can be successfully registered against the transformed watermarked image. In these cases, the watermark detector unambiguously identifies the owner. Further, the use of Gaussian noise, ensures strong resilience to multiple-document, or collusional, attacks. Experimental results are provided to support these claims, along with an exposition of pending open problems.

Asymmetric Fingerprinting For Larger Collusions- AUTHORS: B. Pfitzmann and M. Waidner

Fingerprinting schemes deter people from illegal copying of digital data by enabling the merchant of the data to identify the original buyer of a copy that was redistributed illegally. All known fingerprinting schemes are symmetric in the following sense: Both the buyer and the merchant know the fingerprinted copy. Thus, when the merchant finds this copy somewhere, there is no proof that it was the buyer who put it there, and not the merchant. They introduce asymmetric fingerprinting.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Where only the buyer knows the fingerprinted copy, and the merchant, upon finding it somewhere, can find out and prove to third parties whose copy it was. We present a detailed definition of this concept and constructions. The first construction is based on a quite general symmetric fingerprinting scheme and general cryptographic primitives; it is provably secure if all these underlying schemes are. We also present more specific and more efficient constructions.

A Digital Signature Scheme Secure Against Adaptive chosen-message attacks- AUTHORS: S. Goldwasser, S. Micali, and R. L. Rivest

We present a digital signature scheme based on the computational difficulty of integer factorization. The scheme possesses the novel property of being robust against an adaptive chosen-message attack: an adversary who receives signatures for messages of his choice (where each message may be chosen in a way that depends on the signatures of previously chosen messages) cannot later forge the signature of even a single additional message. This may be somewhat surprising, since in the folklore the properties of having forgery being equivalent to factoring and being invulnerable to an adaptive chosen-message attack were considered to be contradictory. More generally, we show how to construct a signature scheme with such properties based on the existence of a "claw-free" pair of permutations--a potentially weaker assumption than the intractibility of integer factorization.The new scheme is potentially practical: signing and verifying signatures are reasonably fast, and signatures are compact.

A Computational Model For Watermark Robustness-

AUTHORS: A. Adelsbach, S. Katzenbeisser, and A.-R. Sadeghi

Multimedia security strategies generally combine cryptographic strategies with data hiding procedures like steganography or watermarking. Example appliances are dispute resolving, proof of ownership, (asymmetric/anonymous) fingerprinting and zero-knowledge watermark detection.

The need for formal security definitions of watermarking schemes is manifold, whereby the core need is to provide suitable abstractions to develop, analyse and prove the security of applications on top of watermarking schemes. Although there exist formal models and definitions for information-theoretic and computational security of cryptographic and steganographic schemes, they cannot simply be adapted to watermarking schemes due to the fundamental differences among these approaches. Moreover, the existing formal definitions for watermark security still suffer from conceptual deficiencies.

EXISTING SYSTEM:

In the digital era, information leakage through unintentional exposures, or intentional sabotage by disgruntled employees and malicious external entities, present one of the most serious threats to organizations. Confidential data is undoubtedly one of the most severe security threats that organizations face in the digital era. The threat now extends to our personal lives: a plethora of personal information is available to social networks and smart phone providers and is indirectly transferred to untrustworthy third party and fourth party applications.

Encryption process will convert the original information or plain text to cipher text or unidentifiable text and Decryption process is used to get back the original plaintext from the ciphertext. several algorithms are present to carry out these processes and each algorithm will give a different technique to carry out these processes. But using the same algorithm for all data will arise many problems. For example if we use only aes algorithm to convert plain text to cipher text then the attack can get the data by cracking the single algorithm.

Disadvantages:

1. Duplicate data increased.
2. Data leakage is more.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

PROPOSED SYSTEM:

Identification of the leaker is made possible by forensic techniques, but these are usually expensive and don't always generate the desired results. Therefore, we point out the need for a general accountability mechanism in data transfers. This accountability can be directly associated with provably detecting a transmission history of data across multiple entities starting from its origin. This is known as data provenance, data lineage or source tracing. The data provenance methodology, in the form of robust watermarking techniques or adding fake data, has already been suggested in the literature and employed by some industries. However, most efforts have been ad-hoc in nature and there is no formal model available.

Additionally, most of these approaches only allow identification of the leaker in a non-provable manner, which is not sufficient in many cases. We present a generic data lineage framework LIME for data flow across multiple entities that take two characteristic, principal roles (i.e., owner and consumer). We define the exact security guarantees required by such a data lineage mechanism toward identification of a guilty entity, and identify the simplifying non-repudiation and honesty assumptions. We then develop and analyze a novel accountable data transfer protocol between two entities within a malicious environment by building upon oblivious transfer, robust watermarking, and signature primitives.

To encrypt and decrypt the information several algorithms are present. so, the person who want to steal the information should know about that algorithms. If we use single algorithm to encrypt and decrypt the full information he can steal the information easily, but if we use combination of algorithms then that person should know about all the algorithms we have used. so the time needed to crack the information is increased and then security provided to the information is also increased. For example if we are uploading text and an image in the web, we can use AES algorithm for encrypting the text and RSA algorithm for encrypting the image. We can use the advanced AES algorithm by using Padding mechanisms like PKCS5 or PKCS7 and another algorithm to convert the bits output to string. In this way, we can make the hacker confuse about cracking or stealing the information and we can increase the security of our information. Padding mechanisms are used to pad the bits and make the traffic analysis harder for the hacker.

Advantages:

1. We can detect the data leakages.

MODULES:

1. Lime
2. Dataowner
3. Consumer
4. Auditor

Module Description:

LIME:

A generic data lineage framework for dataflow across multiple entities in themalicious environment. we identify an optional non-repudiation assumption made between two owners,and an optional trust (honesty) assumption made by the auditor aboutthe owners.The key advantage of our model is that it enforces accountability by design;

Data owner:

The data owner is responsible for themanagement of documents and the consumer receives documents and can carry out some task using them.

Consumer:

which receives the document. Consumersmight transfer a document to another consumer, sowe also have to consider the case of an untrusted sender. Each consumer can reveal new embedded information to the auditor to point to the next consumer and to prove his own innocence.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Auditor:

which is is not involved in the transfer of documents, it is only invoked when a leakage occurs and then performs all steps that are necessary to identify the leaker

Algorithm

Cipher(byte in[16],byte out[16], key_array round_key [nr+1])

begin

byte state[16]; state = in;

Add Round Key(state, round_key[0]);

for i = 1 to nr-1 stepsize 1 do

Sub Bytes (state);

Shift Rows (state);

Mix Columns (state);

Add Round Key (state,round_key[i]);

End for

Sub Bytes (state);

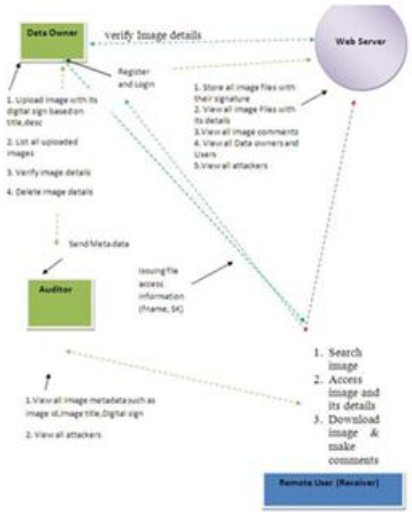
Shift Rows (state);

Add Round Key(state, round_key[nr]);

End

Description:

For example if we take an input(input) of 128 bits it will be converted into 16 bytes and it will undergo all the operations of aes encryption process and they are XORED with 128 bit round key. If this is last round then the output(out) is cipher text otherwise the process will continue and there is no limit for number of rounds(Nr).In the decryption process, the operations are applied reversly on the cipher text and the cipher text is input(in or state) and the output is generated called plain text(out).



Owner Registration Page:

This is the owner registration page where This page is for owner login where owner should enter user name and password and prss submit button. In this page we can see clearly the MENU in right siad

Owner Login page:

This page is for owner login where owner should enter user name and password and prss submit button. In this page we can see clearly the MENU in right said,It has (Home,Owner Consumer…..etc) If we press any one of them will display forexample how many owners are there

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Main Page:

After login web server login page will show view all owner , and all consumer, as well as imaes files. In addition in this page will view all attackers and transactions .Consumer search history and logout options.

View All Owner Page:

In this page will view all owners pages and will display them details (for all owners) Represented them Id , User finger print, User name , Mobile number , Address, Status……..etc.

Admin Page:

Admin page constis of my profile details with, uplod image, list of all uplodedimage, verify owner image files, comments on my images, my imagefile transactions and logout options.

Auditor Login Page:

Auditor login page consits of user name and password options. This is useful for loging in to auditor module.

Auditor Main Page:

Auditor main page is useful for viewing all images files and view all attackers details with logout options. He will look after if any data is attacked by other users.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

All Attacker Page:

In this page list of attackers details are displayed with id ,titlle , name, attacker, attacked data and digital sign.

Register Successful Page:

Registered successful page will be displayed after login registration is successful. In this page will see successful message as well as the back button if we want to go back.

Conclusion:

In this paper, we implement LIME, a model for accountable data transfer across multiple entities. We define participating parties, their interrelationships and give a concrete instantiation for a data transfer protocol using a novel combination of oblivious transfer, robust watermarking and digital signatures. We prove its correctness and show that it is realizable by giving micro bench marking results. By presenting a general applicable framework, we introduce accountability as early as in the design phase of a data transfer infrastructure.

Although LIME does not actively prevent data leakage, it introduces reactive accountability. Thus, it will deter malicious parties from leaking private documents and will encourage honest (but careless) parties to provide the required protection for sensitive data. LIME is flexible as we differentiate between trusted senders (usually owners) and untrusted senders (usually consumers). In the case of the trusted sender, a very simple protocol with little overhead is possible.

The untrusted sender requires a more complicated protocol, but the results are not based on trust assumptions and therefore they should be able to convince a neutral entity (e.g. a judge). Our work also motivates further research on data leakage detection techniques for various document types and scenarios. For example, it will be an interesting future research direction to design a verifiable lineage protocol for derived data.

References:

[1]"Chronology of data breaches," http://www.privacyrights.org/data breach.

[2]"Data breach cost", http//www.symantec.com/about/news/releas/article.jsp ?prid=20110308 01.

[3]"Privacy rights clearinghouse," http://www.privacyrights.org.

[4]"Electronic Privacy Information Center (EPIC)," http://epic.org, 1994.

[5]"Facebook in Privacy Breach," http://online.wsj.com/article/SB10001424052702304772804575558484075236968.html.

[6]"Offshore outsourcing, " http:/www.computerworld.com/article/100938/offshore outsourcing cited in Florida leak.

[7]A. Mascher-Kampfer, H. St ¨ogner, and A. Uhl, "Multiple re-watermarking scenarios," in Proceedings of the 13th International Conference on Systems, Signals, and Image Processing (IWSSIP 2006).Citeseer, 2006, pp. 53–56.

[8]P.Papadimitriou and H. Garcia-Molina, "Data leakage detection," Knowledge and Data Engineering, IEEE Transactions on, vol. 23, no. 1, pp. 51–63, 2011.

[9]"Pairing- Based cryptography Library (PBC)" http://crypto.stanford.pbc

[10]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.