International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 04 Issue 14
November 2017

A Survey Paper on Data Lineage in Malicious Environments

1. **MEHABUNNISA**

2. **MS.K.NAGALATHA**

1.Pg Scholar, Department Of ECE, Annamacharya Institute Of Technology And Sciences,Piglipur, Batasingaram(V), Hayathnagar(M), Ranga Reddy(D),Hyderabad.

2. Asst.Professor and Head of the Department, Department Of ECE, Annamacharya Institute Of Technology And Sciences,Piglipur, Batasingaram(V), Hayathnagar(M), Ranga Reddy(D),Hyderabad

ABSTRACT:

Intentional or unintentional leakage of confidential data is undoubtedly one of the most severe security threats thatorganizations face in the digital era. The threat now extends to our personal lives: a plethora of personal information is available tosocial networks and smartphone providers and is indirectly transferred to untrustworthy third party and fourth party applications. In thiswork, 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 identificationof a guilty entity, and identify the simplifying non-repudiation and honesty assumptions. We then develop and analyze a novelaccountable data transfer protocol between two entities within a malicious environment by building upon oblivious transfer, robustwatermarking, and signature primitives. Finally, we perform an experimental evaluation to demonstrate the practicality of our protocoland apply our framework to the important data leakage scenarios of data outsourcing and social networks. In general, we consider LIME, our lineage framework for data transfer, to be an key step towards achieving accountability by design.

1 INTRODUCTION

IN the digital era, information leakage through unintentionalexposures, or intentional sabotage by disgruntledemployees and malicious external entities, present one ofthe most serious threats to organizations. According to aninteresting chronology of data breaches maintained by thePrivacy Rights Clearinghouse (PRC), in the United Statesalone, 868;045;823 records have been breached from 4;355data breaches made public since 2005 [1]. It is not hard tobelieve that this is just the tip of the iceberg, as most casesof information leakage go unreported due to fear of loss ofcustomer confidence or regulatory penalties: it costs companieson average $214 per compromised record [2]. Largeamounts of digital data can be copied at almost no cost andcan be spread through the internet in very short time. Additionally,the risk of getting caught for data leakage is verylow, as there are currently almost no accountability mechanisms.For these reasons, the problem of data leakage hasreached a new dimension nowadays.Not only companies are affected by data leakage, it isalso a concern to individuals. The rise of social networksand smartphones has made the situation worse. In theseenvironments, individuals disclose their personal informationto various service providers, commonly known as thirdparty applications, in return for some possibly free services.In the absence of proper regulations and accountabilitymechanisms, many of these applications share individuals'identifying information with dozens of advertising andInternet tracking companies.Even with access control mechanisms, where access tosensitive data is limited, a malicious authorized user canpublish sensitive data as soon as he receives it. Primitiveslike encryption offer protection only as long as the informationof interest is encrypted, but once the recipient decryptsa message, nothing can prevent him from publishing thedecrypted content. Thus it seems impossible to prevent dataleakage proactively.Privacy, consumer rights, and advocacy organizationssuch as PRC [3] and EPIC [4] try to address the problem ofinformation leakages through policies and awareness. However,as seen in the following scenarios the effectiveness ofpolicies is questionable as long as it is not possible to provablyassociate the guilty parties to the leakages.

2 RELATED WORK

A preliminary shorter version of this paper appeared at the STM workshop . This version constitutes a significantextension by including the following contributions:We give a more detailed description of our model, a formalspecification of the used primitives, an analysis ofthe introduced protocol, a discussion of implementationresults, an application of our framework to examplescenarios, a discussion of additional features and anextended discussion of related work.Clustering analysis is veryuseful to estimate the inter-entity similarity. One good example

of clustering based reranking algorithms is the InformationBottle based scheme developed by Hsu et al.[9]. In thismethod, the images in the initial results are primarily groupedautomatically into several clusters. Then the re-ranked resultlist is created first by ordering the clusters according tothe cluster conditional probability and next by ordering thesamples within a cluster based on their cluster membership value. In a fast and accurate scheme is proposed forgrouping Web image search results into semantic clusters. Itis obvious that the clustering based reranking methods canwork well when the initial search results contain many nearduplicate media documents. However, for queries that returnhighly diverse results or without clear visual patterns, theperformance is not guaranteed.

3 THE LIME FRAMEWORK

As we want to address a general case of data leakage in datatransfer settings, we propose the simplifying model LIME(Lineage in the malicious environment). With LIME weassign a clearly defined role to each involved party anddefine the inter-relationships between these roles. Thisallows us to define the exact properties that our transferprotocol has to fulfill in order to allow a provable identificationof the guilty party in case of data leakage.

3.1 Model

As LIME is a general model and should be applicable to allcases, we abstract the data type and call every data item document.There are three different roles that can be assigned tothe involved parties in LIME: data owner, data consumer andauditor. The data owner is responsible for the managementof documents and the consumer receives documents andcan carry out some task using them. The auditor is notinvolved in the transfer of documents, he is only invokedwhen a leakage occurs and then performs all steps that arenecessary to identify the leaker. All of the mentioned rolescan have multiple instantiations when our model is appliedto a concrete setting. We refer to a concrete instantiation ofour model as scenario.In typical scenarios the owner transfers documents toconsumers. However, it is also possible that consumers passon documents to other consumers or that owners exchangedocuments with each other.

In the outsourcing scenario [6]the employees and their employer are owners, while theoutsourcing companies are untrusted consumers.In the following we show relations between the differententities and introduce optional trust assumptions. We onlyuse these trust assumptions because we find that they arerealistic in a real world scenario and because it allows us tohave a more efficient data transfer in our framework. At theend of this section we explain how our framework can beapplied without any trust assumptions.When documents are transferred from one owner toanother one, we can assume that the transfer is governed bya non-repudiation assumption. This means that the sendingowner trusts the receiving owner to take responsibility ifhe should leak the document. As we consider consumersas untrusted participants in our model, a transfer involvinga consumer cannot be based on a non-repudiation assumption.Therefore, whenever a document is transferred to aconsumer, the sender embeds information that uniquelyidentifies the recipient. We call this fingerprinting. If the consumerleaks this document, it is possible to identify himwith the help of the embedded information.As presented, LIME relies on a technique for embeddingidentifiers into documents, as this provides an instrumentto identify consumers that are responsible for data leakage.We require that the embedding does not not affect the utilityof the document. Furthermore, it should not be possiblefor a malicious consumer to remove the embedded informationwithout rendering the document useless. A techniquethat can offer these properties is robust watermarking. Wegive a definition of watermarking and a detailed descriptionof the desired..

4 ACCOUNTABLE DATA TRANSFER

In this section we specify how one party transfers a documentto another one, what information is embedded andwhich steps the auditor performs to find the guilty party incase of data leakage. We assume a public key infrastructureto be present, i.e., both parties know each others signatureverification key.

4.1 Trusted Sender

In the case of a trusted sender it is sufficient for the sender toembed identifying information, so that the guilty party canbe found. As the sender is trusted, there is no need for furthersecurity mechanisms. we present a transferprotocol that fulfills the properties of correctness and nodenial as. As the sender is trusted tobe honest, we do not need the no framing property.The sender, who is in possession of some document D,creates a watermarking key k, embeds a triple consisting of the two parties' identifiers and a timestampt into D to create Dw ¼WðD; s; kÞ. He then sends Dwto the recipient, who will be held accountable for thisversion of the document. As the sender also knows Dw, thisvery simple protocol is only applicable if the sender iscompletely trusted; otherwise the sender could publish Dwand blame the recipient.

4.2 Untrusted Sender

In the case of an untrusted sender we have to take additionalactions to prevent the sender from cheating, i.e., wehave to fulfill the no framing property. To achieve this property,the sender divides the original document into n partsand for each part he creates two differently watermarkedversions. He then transfers one of each of these two versionsto the recipient via OT2

1 . The recipient is held accountableonly for the document with the parts that he received, butthe sender does not know which versions that are. Theprobability for the sender to cheat is therefore 12n. We showthe protocol and provide an analysis of the protocolproperties.First, the sender generates two watermarking keys k1 andk2. It is in his own interest that these keys are fresh and distinct.The identifying information that the sender embedsinto the documentD is a signed statement s ¼ ½CS; CR; t_skCRcontaining the sender's and recipient's identifiers and atimestamp t, so that every valid watermark is authorized bythe recipient. The sender computes the watermarked documentsplits the document D0 into n partsand creates two different versions

4.3 Data Lineage Generation

The auditor is the entity that is used to find the guilty partyin case of a leakage. He is invoked by the owner of the documentand is provided with the leaked document. In order toProtocol for trusted senders: The sender watermarks the originaldocument with a signed statment containing the participants' identifiersand a timestamp, and sends the watermarked document to the recipient. find the guilty party, the auditor proceeds in the followingway:

1) The auditor initially takes the owner as the currentsuspect.

2) The auditor appends the current suspect to thelineage.

3) The auditor sends the leaked document to the currentsuspect and asks him to provide the detectionkeys k1 and k2 for the watermarks in this documentas well as the watermark s. If a non-blind watermarkingscheme is used, the auditor additionallyrequests the unmarked version of the document.

4) If, with key k1, s cannot be detected, the auditor continueswith 9.

5) If the current suspect is trusted, the auditor checksthat s is of the form where CS is the identifierof the current suspect, takes CR as current suspectand continues with 2.

6) The auditor verifies that s is of the form ½CS;CR; t_skCRwhere CS is the identifier of the currentsuspect. He also verifies the validity of the signature.

7) The auditor splits the document into n parts and foreach part he tries to detect 0 and 1 with key k2. Ifnone of these or both of these are detectable, he continueswith 9. Otherwise he sets b0i as the detected bitfor the ith part. He sets b0 ¼ b01 . . . b0n.

8) The auditor asks CR to prove his choice of b ¼ b1 _ _ _ bn for the given timestamp t by presenting the. If CR is not able to give a correctproof (i.e., mi;bi is of the wrong form or the signatureis invalid) or if b ¼ b0, then the auditor takes CR ascurrent suspect and continues with 2.

9) The auditor outputs the lineage. The last entry isresponsible for the leakage.

CONCLUSION AND FUTURE DIRECTIONS

We present LIME, a model for accountable data transferacross multiple entities. We define participating parties,their inter-relationships and give a concrete instantiation fora data transfer protocol using a novel combination of oblivioustransfer, robust watermarking and digital signatures.We prove its correctness and show that it is realizable bygiving microbenchmarking results. By presenting a generalapplicable framework, we introduce accountability as earlyas in the design phase of a data transfer infrastructure.Although LIME does not actively prevent data leakage, itintroduces reactive accountability. Thus, it will deter maliciousparties from leaking private documents and willencourage honest (but careless) parties to provide therequired protection for sensitive data. LIME is flexible as wedifferentiate between trusted senders (usually owners) anduntrusted senders (usually consumers). In the case of thetrusted sender, a very simple protocol with little overheadis possible. The untrusted sender requires a more complicatedprotocol, but the results are not based on trustassumptions and therefore they should be able to convincea neutral entity (e.g., a judge).Our work also motivates further research on dataleakage detection techniques for various document typesand scenarios. For example, it will be an interestingfuture research direction to design a verifiable lineageprotocol for derived data.

.REFERENCES

[1] Chronology of data breaches [Online]. Available: http://www.privacyrights.org/data-breach, 2014.

[2] Data breach cost [Online]. Available: http://www.symantec.com/about/news/release/article.jsp?prid=20110308_01, 2011.

[3] Privacy rights clearinghouse [Online]. Available: http://www.privacyrights.org, 2014.

[4] (1994). Electronic privacy information center (EPIC) [Online].Available: http://epic.org, 1994.

[5] Facebook in privacy breach [Online]. Available: http://online.wsj.com/article/SB10001424052702304772804575558484075236968.html, 2010.

[6] Offshore outsourcing [Online]. Available: http://www.computerworld.com/s/article/109938/Offshore_outsourcing_cited_in_Florida_data_leak, 2006.

[7] A. Mascher-Kampfer, H. St€ogner, and A. Uhl, “Multiple re-watermarkingscenarios,” in Proc. 13th Int. Conf. Syst., Signals, ImageProcess., 2006, pp. 53–56.

[8] P. Papadimitriou and H. Garcia-Molina, “Data leakage detection,”IEEE Trans. Knowl. Data Eng., vol. 23, no. 1, pp. 51–63, Jan. 2011.

[9] Pairing-based cryptography library (PBC) [Online]. Available:http://crypto.stanford.edu/pbc, 2014.

[10] I. J. Cox, J. Kilian, F. T. Leighton, and T. Shamoon, “Secure spreadspectrum watermarking for multimedia,” IEEE Trans. ImageProcess., vol. 6, no. 12, pp. 1673–1687, Dec. 1997.

[11] B. Pfitzmann and M. Waidner, “Asymmetric fingerprintingfor larger collusions,” in Proc. 4th ACM Conf.

Comput. Commun.Security, 1997, pp. 151–160.

[12] S. Goldwasser, S. Micali, and R. L. Rivest, “A digital signaturescheme secure against adaptive chosen-message attacks,” SIAMJ. Comput., vol. 17, no. 2, pp. 281–308, 1988.

[13] A. Adelsbach, S. Katzenbeisser, and A.-R. Sadeghi, “A computationalmodel for watermark robustness,” in Proc. 8th Int. Conf. Inf.Hiding, 2007, pp. 145–160.

[14] J. Kilian, F. T. Leighton, L. R. Matheson, T. G. Shamoon, R. E.Tarjan, and F. Zane, “Resistance of digital watermarks tocollusive attacks,” in Proc. IEEE Int. Symp. Inf. Theory, 1998,pp. 271–271.

[15] M. Naor and B. Pinkas, “Efficient oblivious transfer protocols,” inProc. 12th Annu. ACM-SIAM Symp. Discrete Algorithms, 2001,pp. 448–457.

[16] GNU multiple precision arithmetic library (GMP) [Online]. Available:http://gmplib.org/, 2014.

[17] D. Boneh, B. Lynn, and H. Shacham, “Short signatures fromthe Weil pairing,” in Proc. 7th Int. Conf. Theory Appl. Cryptol. Inf.Security: Adv. Cryptol., 2001, pp. 514–532.

[18] W. Dai. Crypto++ Library [Online]. Available: http://cryptopp.com, 2013.

[19] P. Meerwald. Watermarking toolbox [Online]. Available: http://www.cosy.sbg.ac.at/ pmeerw/Watermarking/source, 2010.

[20] Y. Ishai, J. Kilian, K. Nissim, and E. Petrank, “Extending oblivioustransfers efficiently,” in Proc. 23rd Annu. Int. Cryptol. Conf. Adv.Cryptol., 2003, pp. 145–161.