Efficient Secure Provenance Scheme for Strong Integrity

Myat Mon Aye, and Than Naing Soe

***Abstract*—**As escalating the amount of precious information are recorded digitally especially on cloud storage and web, the ability to detect the ownership, responsibility and history of the specific item plays the vital role for management activities such as rights protection, digital forensic, regulatory compliance. The most promising technology is data provenance which maintains the chronology of the ownership and process history. Although the significant researches have been carried out in this area, very little researches have been conducted for security and privacy issues of provenance data. Moreover, the previous researches for security issues have defect on integrity and overhead. In this paper, new secure provenance scheme is proposed to provide strong integrity with reasonable overhead. This paper shows how to support four dimensional security requirements called confidentiality, privacy, integrity and availability. This scheme also provides the privacy (confidentiality) for user's sensitive data using broadcast encryption with divisive clustering algorithm.

***Keywords*—**Broadcast encryption, hierarchical clustering provenance, secure provenance.

I. Introduction

PROVENANCE is derived from French word "provenir" which means origin and ownership of an historical item. However, its sense is more meaningful in this era because provenance refers the ownership, process activities, workflow and lineage information of an object and it is vital for all kinds of organization such as government offices, commerce, science etc...

As an example, today, military uses computerized system to maintain the official information, orders and records of activities. Alice give order to Bob without permission of senior officer, and then Bob forwards this order to Charlie to perform operation. When the senior officers know about that operation, the main important challenge is to decide who is a responsible person Alice, Bob or Charlie. In this situation, if the provenance of orders and workflow isn't recorded systematically, it leads to the collusion.

Provenance chain tracking is very important for organizations as they record and carry out their new findings, projects, technology and manufacturing process digitally. Due to the very valuable information are recorded, the organization must strongly determine who has authority to access these information. However, adversaries, from outside or inside of the organization, try to access the precious and leak to competitors. At this time, the authorized person in the organization can be bottleneck to prove they aren't responsible if the outside adversaries leak this precious information. In this situation, the authorized person can have a big problem to reveal the offender. If the organization tacks and records the access history (provenance), it is not need to effort to prove innocent.

In this condition, provenance information is vital information itself so the security of provenance information is an important issue and the provenance chain is also important to reveal the time oriented activities. This paper presents the efficient secure provenance scheme using cryptographic primitives and protocols. The main contributions of this paper are (1) propose a new scheme of secure provenance with low overhead, (2) use time-stamp oriented provenance chain for strong integrity and (3) propose hybrid grouping method in Broadcast Encryption for reducing overhead of public key distribution.

This paper is organized as follows. Section 2 describes the related work. The background theory of secure provenance is discussed in Section 3. In Section 4, secure provenance scheme is proposed and this paper is then concluded in Section 5.

II. Related Work

Provenance has been observed extensively for various kinds of areas to track the origin, activities and access history. Kiran-Kumar Muniswamy-Reddy, Peter Macko, and Margo Seltzer [3] designed and implemented three protocols to record the provenance data for cloud stores. They also evaluated the proposed three protocols to point out the reasonable overhead and individual properties.

In [4], Masoud Valafar and Kevin Butler proposed a model for secure provenance collection, storage and management for cloud storage. In this, Provenance Aware Storage System (PASS) was applied for collection and storage and PQL, provenance query language, was used for auditors. They also implemented their proposed model with python over Cumulus, the storage system for Nimbus open source cloud toolkit and presented the analysis of performance over various file size.

Myat Mon Aye is with University of Computer Studies, Mandalay, Myanmar, (corresponding e-mail: myatmonaye.j18@gmail.com).

Than Naing Soe was with the Department of Hardware, University of Computer Studies, Mandalay, Myanmar. He is now with Computer University, Myit Kyi Na, (e-mail: konaing2006@gmail.com).

Imad M. Abbadi and Johm Lyle [2] demonstrated the challenges of provenance collection in cloud computing. In [1], Boris Glavic and Klaus Dittrich surveyed data provenance models and schemes and presented the categorization scheme for existing approaches.

Hasan and others, [5] described the main challenges in trustworthy of provenance and defined an adversarial model and analyzed the potential security and privacy issues related to securing provenance information.

Hansan Rigib and Winslett Marianne [6] implemented secure provenance scheme at application layer for file system by handing the message digest of preceding provenance records in check sum and produced the experimental results for time overhead with (i) Postmark- a standard benchmark for file system performance evaluation (ii) small and large file microbenchmarks that have been used to evaluate the performance of PASS and (iii) a custom transactional benchmark.

Xinlei (Oscar) Wang and others [8] proposed a chain structure provenance scheme to sure the three dimensions of security called confidentiality, integrity and availability by using a public key linked chain. Their work mainly focused on two parts: adding own provenance record and verifying provenance chain. They also provided the empirical assessments for the overhead of provenance construction, verification and storage.

III. Background Theory

In this section, what is provenance and its application areas, the important of provenance data security and the main challenges of secure provenance are discussed.

A. Provenance

Provenance, from the French word provenir, "to come from", refers to the chronology of the ownership or location of a historical object [10]. It is also called information lineage and important for digital forensics and post-incident because it can stores the ownership and process history of specific object.

For example, business organizations carry out their tasks via online in order to accomplish easily and quickly. The management of the specific organization desires to monitor the facts such as who are the responsible person for a particular file, who accessed to the important files and what did on it. The provenance can fulfill all of these requirements by storing each creation or access as a provenance record and then regards the time order sequence of provenance records as provenance chain. However, adversaries, who can be inside or outside of the organization, want to access the provenance record and alter the order of provenance chain or the contents of a particular provenance record in order to hide the unauthorized manners. To solve this above problem, the security is an essential requirement for the provenance.

The main research areas relating with data provenance are provenance collection, provenance representation, query, and storage and security issue. Existing studies of provenance management mainly focused on the collection, representation, query and storage of provenance data [7].

B. Secure Provenance

The secure provenance issue can be addressed via the tasks of providing assurances of confidentiality (privacy), integrity and availability.

Confidentiality and Privacy: A provenance record may contain information that is of a confidential nature in two ways:

1) Information about the tasks performed may be secret;

2) The ownership history might contain sensitive information that should not be revealed to unauthorized parties.

Integrity: The integrity of provenance is three-fold:

1) Data Integrity: The information contained in each individual provenance record is not tampered with.

2) Origin Integrity: The origin of each individual provenance record is not forged. A node cannot deny its ownership of the provenance records created by it.

3) Chain Integrity: The order of the owners on the provenance chain can't be modified [8].

Availability: Availability must ensure that no provenance records in the provenance chain can be selectively dropped without being detected object.

IV. Efficient Secure Provenance Scheme

The proposed efficient provenance scheme is explained with two sections such as system model and chain construction and security issue. In system model, the three main principals and their responsibilities are introduced. The details of chain construction and how to solve security issues are discussed in the next section.

A. System Model

In the proposed scheme, there are three main principals called System Manager (SM), Key Distributor (KD) and Chain of Users {U1, U2 … Un} as shown in figure 1.

1) System Manager (SM): System manager is a trustable and powerful entity. The responsibility of SM is in charge of the management of the whole system and serves as a witness when the new user access or derived the existing document.

2) Key Distributor (KD): Key distributor is a responsible person for all activities of key management such as key generation, key distribution.

3) Users (Ui): Users are principals who read and write document and their metadata. The users can be further classified into three types such as:

a) Auditor- is an authorized person to verify the integrity of provenance records associated with document.

b) Super auditor- is the most authorized person to access and check the provenance chain and all details of the provenance record.

c) Adversaries- are principals from inside or outside of the organization who access to a document and it provenance chain and want to alter them inappropriately with devious purposes.

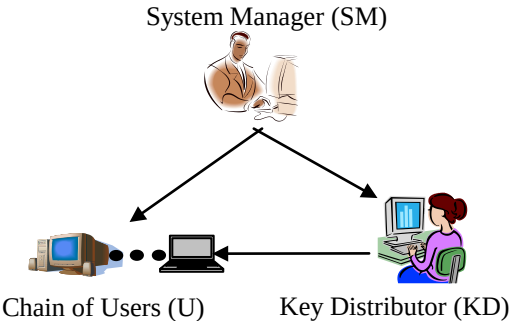

Fig. 1 System Model for Efficient Secure Provenance Scheme

B. Secure Provenance

In this section, the steps of chain construction and supporting security issues are explained. A provenance chain for a given document is constructed with the time-ordered sequence of provenance records P1|P2 • • • |Pi| • • • |Pn| and each provenance record Pi contains eight materials called Ti, Ui, Ai, hash(Douti), PKCi, Ikeysi, Ci and ITokeni as in Fig 2.

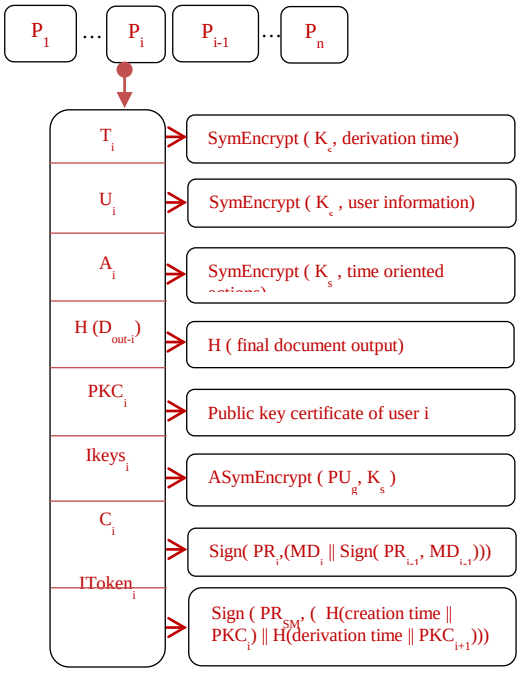

Fig. 2 Overall Structure of Provenance Chain

In the provenance record, Ti is opaque form of creation time of Pi or inheritance time of Pi-1 e.g. Fri Dec 07 08:58:28 MMT 2012, Ui is opaque form of user information e.g. User ID, IP Address etc., Ai is opaque form of time oriented user actions e.g. Read; Write, Copy, Delete etc., hash (Douti) is message digest of the final output document from user actions, PKCi is the public key certificate of user i, Ikeysi is all key materials of interpreting the preceding fields, Ci is integrity checksum for preceding record and ITokeni is the collections of access time and public key certificate of all inherits for integrity purpose.

The pseudo code of the provenance chain construction is illustrated in figure 3.

1. Set dTime to derivation time
2. Set u to user identity information
3. Set a to time oriented actions performed by user i
4. Set PKCi to public key certificate of user i
5. Set Douti to output document of user i
6. Set cTime to creation time
7. Ti = SymEncrypt (Ks, dTime)
8. Ui = SymEncrypt (Ks, u)
9. Ai = SymEncrypt (Ks, a)
10. HDi = H (Douti)
11. for each g ε Accessible groups
    Ikeysi = Ikeysi ∪ ASymEncrypt (PUg, Ks)
    Next g
12. MDi = H (Ti, Ui, Ai, HDi, PKCi, Ikeysi)
13. Ci = Sign (PRi, (MDi || Sign (PRi-1,MDi-1)))
14. ITokeni = Sign(PRSM (H (cTime || PKCi)))
15. Pi = <Ti, Ui, Ai, hash (Douti), PKCi, Ikeysi, Ci, ITokeni>
16. if (Di is derived) then
    ITokeni = Sign (PRSM ,(H (cTime ||PKCi) || H(dTime|| PKCi+1)))
    End if

Fig. 3 Pseudo code for provenance chain construction

Where, H ( ) is hash function, SymEncrypt is symmetric encryption, ASymEncrypt is asymmetric encryption, Ks is secret key, PRi and PKCi are private key and public key certificate of user i, PRSM is private key of system manager and PUg is public key of group g.

According to these aforementioned descriptions, each provenance record consists of eight materials. Since the first three materials contain the sensitive information, this information is transformed into opaque form in order to ensure confidentiality and privacy issue. In this proposed system, opaque form is provided using the broadcast encryption with hybrid grouping method in order to reduce the overhead of key distribution for efficiency.

In broadcast encryption, the users are clustered based on the previous sharing patterns, so it is suitable with hierarchical clustering algorithm because some document are shared to specific user, some for all user and some for specific users group. Among two popular hierarchical clustering algorithms called Agglomerative clustering (bottom up) and Divisive Clustering (top down), Divisive clustering is used because it takes less complexity than Agglomerative clustering [9].

The procedure of Divisive clustering algorithm is as follow:

1. Put all objects in one cluster.
2. Repeat until all clusters are singletons
   a. Choose a cluster to split.
   b. Replace the chosen cluster with the sub-clusters.

After clustering, the KD produces the private and public key

pair to each group as in following figure.

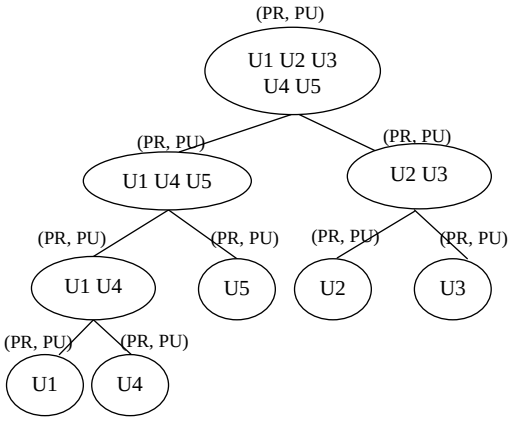

Fig. 4 Groups and key pairs

The user sensitive information is encrypted using the session key and then this session key is encrypted using the public key of the intended groups and the resulted cipher text is inserted into Ikeys. According clustered tree, the user can encrypt the session key with top node's public key if the user wants to allow all users to access sensitive data.

$$Ti = SymEncrypt (Ks, derivation time) \quad (1)$$ (1)

$$Ui = SymEncrypt (Ks, user information) \quad (2)$$ (2)

$$Ai = SymEncrypt (Ks, time oriented actions) \quad (3)$$ (3)

For each group g,

$$Ikeys = Ikeys \cup ASymEncrypt (PUg, Ks) \quad (4)$$ (4)

When an adversary wants to check the sensitive information Ikeys is firstly decrypts with its private key or group's private key for session key which is used to reveal the sensitive information.

Data integrity is achieved by comparing the message digest of the output document of previous user Pi-1 with the hash code of current document to user i using forth material in the provenance record.

The last two materials called C and IToken (Integrity Token) are constructed for provenance chain integrity. C is created to verify the integrity of the provenance chain from the beginning to the end by handing the message digest of vital information of previous record.

$$MDi=H(Ti, Ui, Ai, H (Douti), PKCi, Ikeysi) \quad (5)$$

$$Ci=Sign (PRi, (MDi \| Sign (PRi-1,MDi-1))) \quad (6)$$

By handing over the previous provenance information with C, the preceding provenances in the chain can't be deleted without being detected.

The creation of IToken is used to link current provenance with the next one.

$$ITokeni = Sign [PRSM , (H (tcreate \| PKCi) \| H(tDrive \| PKCi+1))] \quad (7)$$

Where, PRSM is the private key of SM. By linking the current provenance with the next one via IToken, the last provenances in the chain can't be deleted without being detected.

By ensuring both C and IToken verification, any middle provenance records in the chain can't be deleted without being detected. So, C and IToken construction, for integrity purpose, also provides the availability issue.

V. Conclusion

Security of provenance records is serious requirement in all areas such as business organization, arts, publications, research fields, science, medicine, government and so on for patents, proving authorship and other intellectual property litigations. Moreover, existing secure provenance schemes can't provide strong integrity of provenance chain. This paper proposes a new time oriented secure provenance scheme in order to provide strong integrity with low overhead. According to the revealing of the proposed scheme in previous section, this scheme provides the four security requirements called confidentiality, privacy, integrity and availability. The proposed scheme is implemented based on AES as symmetric encryption, MD5 for hash code, RSA for asymmetric encryption and digital signature and Divisive hierarchical clustering to group users for broadcast encryption to track the file accesses. Experiments point out that the proposed scheme takes the overhead of 1-14% over typical workloads.

References

[1] Boris Glavic, Klaus Dittrich, "Data Provenance: A Categorization of Existing Approaches," NFS SESAM.
[2] Imad M. Abbadi and John Lyle, "Challenges for Provenance in Cloud Computing," USENIX, 2011.
[3] Kiran-Kumar Muniswamy-Reddy, Peter Macko, and Margo Seltzer, "Provenance for the Cloud," USENIX.
[4] Masoud Valafar and Kevin Butler, "Secure Provenance for Cloud Storage," IEEE, 2011.
[5] Ragib Hasan et. al, "Introducing SecureProvenance: Problems and Challenges," ACM, October 29, 2007.
[6] Rigib Hansan and Marianne Winslett, "" Preventing History Forgery with Secure Provenance," ACM Transactions on Storage, Vol. 5, No. 4, Article 12, December 2009.
[7] Shouhuai Xu et. al,, "A Characterization of The Problem of Secure Provenance Management," IEEE, 2009.
[8] Xinlei (Oscar) Wang et. al,, "Chaining for Securing Data Provenance in Distributed Information Networks," IEEE International Conference for Military Communications (MILCOM), 2012.
[9] Ying Zhao and George Karypis, "Comparison of Agglomerative and Partitional Document Clustering Algorithms," Conference on Information and Knowledge Management (CIKM), 2002.
[10] http://en.wikipedia.org/wiki/Provenance