A Characterization of The Problem of Secure Provenance Management

Shouhuai Xu
Department of Computer Science
University of Texas at San Antonio
shxu@cs.utsa.edu

Qun Ni and Elisa Bertino
Department of Computer Science
Purdue University
{ni,bertino}@cs.purdue.edu

Ravi Sandhu
Institute for Cyber Security
University of Texas at San Antonio
ravi.sandu@utsa.edu

***Abstract*—Data (or information) provenance has many important applications. However, prior work on data provenance management almost exclusively focused on the collection, representation, query, and storage of provenance data. In contrast, the security aspect of provenance management has not been understood nor adequately addressed. A natural question then is: What would a secure provenance management system — perhaps as an analogy to secure database management systems — look like? In this paper, we explore the problem space of secure provenance management systems with an emphasis on the *security requirements* for such systems, and characterize desired solutions for tackling the problem. We believe that this paper makes a significant step towards a comprehensive solution to the problem of secure provenance management.**

I. Introduction

What is provenance? The Merriam-Webster Online Dictionary defines provenance as: (1) Origin, source. (2) The history of ownership of a valued object or work of art or literature. The Oxford English Dictionary defines provenance as: (1) The fact of coming from some particular source or quarter; origin, derivation. (2) The history of the ownership of a work of art or an antique, used as a guide to authenticity or quality; a documented record of this. Such notions are very relevant to digital information and have resulted in different kinds of provenance:

- Why-provenance: Why is a certain piece of information here [7], [15]?
- Source-provenance: What is the source of a certain piece of information [7], [15]?
- How-provenance: How does a certain piece of information get here [8]?

Why is provenance important? Data or information provenance has many applications, including: verification of scientific data and experiments [3], [26], [12], [24], [23], [6], supporting/facilitating data sharing [25], [14], [18], [13], copyright clearance [21], legal proceedings involving data [17], information quality (e.g., for determining whether the sources of database tuples in the result of a query are independent [2]).

State-of-the-art in provenance management. Existing studies of provenance management mainly focused on the collection, representation, query and storage of provenance data. For example, the first provenance challenge aimed to understand the expressiveness of provenance representations and capabilities of provenance systems [19]; whereas interoperability is the focus of the Second Provenance Challenge (see http://twiki.ipaw.info/bin/view/Challenge/SecondProvenanceChallenge).

A. Our Contributions

In this paper we make the following contributions:

- We explore the requirements for secure provenance management systems. We argue that secure provenance management systems should cover the whole information lifecycle. In particular, we discuss the security requirements for provenance management systems with an emphasis on the needs of enforcing advanced access control, integrity, accountability, privacy protection, and compliance. We also discuss services that provenance management systems should provide to applications, namely information trustworthiness management, secure information dissemination management, and information compliance management.
- We propose a framework for secure provenance management. The framework accommodates the aforementioned requirements for secure provenance management systems.

Paper organization. Section II discusses the challenges posed by secure provenance management systems. Section III presents a framework for secure provenance management systems. Section IV discusses related prior work. Section V concludes the paper.

II. Requirements for Secure Provenance Management

Secure provenance management is important to many applications, especially in situations where data trustworthiness is a key concern. This naturally inspires us the following question:

> What would secure provenance management systems — say, as an analogy to secure database management systems — look like? How should we design and implement them?

In order to answer the above questions, we first need to understand the problem space of secure provenance management, including the unique challenges posed by it. In this section we explore the requirements and challenges of *secure* provenance management.

A. Functional Requirements

Without loss of generality, we assume that data (or information) may move within distributed/decentralized systems in the format of messages. Moreover, new messages may be produced by algorithms that may take other messages as inputs. That is, we are primarily dealing with data and provenance management in distributed or decentralized systems, which might often be large-scale.

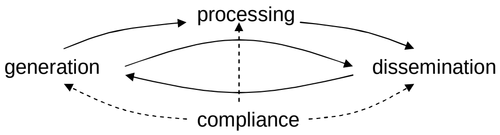

Fig. 1. Information lifecycle in the context of secure provenance management systems (solid lines corresponding to major procedures, and dashed lines indicating compliance applies to the relevant procedures)

From a functional perspective, we believe that a secure provenance management system should cover the whole lifecycle of data as well as their associated provenance. In this context, we classify data lifecycle into the following procedures (see also Figure 1): generation, processing, dissemination, and compliance. We note that this lifecycle is somewhat tailored to secure provenance management systems, and thus may not be appropriate for other systems.

- Generation: A data item originally enters into a provenance management system through some participant; such participant is the party responsible for the initial generation and insertion of the data item into the system.
- Processing: Each participant, source or intermediate node, can produce new data or information items based on the items it received from other participants. Various (e.g., datamining or knowledge extraction) algorithms and functions may be applied to process data items. For example, such a function can simply consists of endorsing a data item another participant is disseminating.
- Dissemination: A participant can disseminate the items it produced to other participants, possibly based on some policies. It is important to ensure that data can be disseminated even in the presence of malicious attackers.
- Compliance: It is important to manage *who could read/write/modify* as well as *who have read/written/modified* which data items. This is especially important, for example, when we need to pin down who are the suspects of information leakage (i.e., identifying insiders).

B. Security Requirements

In this section we elaborate what services should be provided by secure provenance management systems to higher layer applications (Section II-B1), and how data provenance should be secured (Section II-B2).

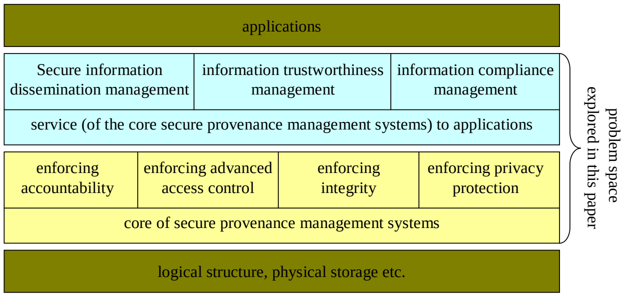
applications
Secure information dissemination management
information trustworthiness management
information compliance management
service (of the core secure provenance management systems) to applications
enforcing accountability
enforcing advanced access control
enforcing integrity
enforcing privacy protection
core of secure provenance management systems
logical structure, physical storage etc.
problem space explored in this paper

Fig. 2. Security requirements of secure provenance management systems

1) Security Service Requirements: A secure provenance management system should provide the following services to higher layer applications: information trustworthiness management, secure information dissemination management, and information compliance management (see also Figure 2).

Information trustworthiness management. In general, information trustworthiness depends on the trustworthiness of the source, the trustworthiness of the intermediate nodes as well as their processing algorithms. However, things quickly become complex when some participants (i.e., sources and intermediate nodes) may be malicious. In what follows we discuss some representative issues relevant to how information trustworthiness should be managed.

- For a source, it is necessary to know about the trustworthiness of a data or information item that has to be entered into the system. It is also necessary that, when a source realizes that it has entered into the system inaccurate or even misleading information (e.g., deceiving information provided by an adversary deliberately), the source be able to inform all the relevant participants about this fact (possibly also to provide updated information).

- For an intermediate node, it is necessary to know about the trustworthiness of both the source and the prior intermediate nodes so that, for example, a decision may be made whether to re-disseminate the (processed) information. It is also important to allow a node to notify upstream nodes (for example) that some data items they provided are inaccurate or even misleading (we call "backward error correction"), and to notify downstream nodes (for example) that some data items they received are inaccurate or even misleading (we call "forward error correction").
- For a data/information consumer, it is necessary to be able to evaluate the trustworthiness of an incoming data/information item. Moreover, the consumer must be cautious in making decisions that rely on such items because the decisions may not be undoable and, once enforced, may cause severe consequences. This may be alleviated using, for example, evidence-based reasoning.
- For an administrator, it is important to know who has a large influence or impact on the evolution of information in the networks. Enhancing security of such participants would significantly improve security from a whole-system perspective.

Secure information dissemination management. The services should address the following questions: What if there are malicious insiders/attackers in the dissemination systems? How can we enforce re-dissemination control? How should the dissemination management help realize the aforementioned backward and forward error correction (even if there are malicious participants who may aim to disrupt the system)? Should information providers passively answer queries with, or proactively push, information according to a policy? When should provenance information be provided (e.g., only when the consumer asks or whenever the provider pushes it?) Is the dissemination process reliable and secure (e.g., conforming to a policy), even in wireless environments under malicious attacks? Is the dissemination process privacy-preserving (e.g., in the procedure of authenticating the information requestors)?

Information compliance management. It is important to answer questions such as:

- Who *has read/written/modified* and *could read/write/modify* a certain data item? This is important because it can help detect, for example, the malicious participant who has leaked a certain confidential data item.
- Who *has read/written/modified* and *could read/write/modify* a certain provenance data item? This is important because it can help detect the party who leaked, for example, "who has participated in which operation/process."

Addressing the above questions will also help resolve the aforementioned problems of backward and forward error correction.

2) Securing Data Provenance: A secure provenance management system should support the following functions: advanced access control, integrity assurance, accountability, and privacy protection (see also Figure 2).

Enforcing advanced access control. The provenance of a data item is often a Directed Acyclic Graph (DAG), where each node represents an object and each arc captures the relationship between two objects [5]. As discussed in [5], traditional access control models do not apply to DAGs, and straightforward adaptations of existing access control models to provenance data do not appear to be sufficient.

Enforcing integrity. Integrity of both data and provenance information is important. Without integrity guarantees, security can be severely undermined.

Enforcing accountability. On one hand, we need to ensure that the participant who maliciously entered false or misleading information is bound to be held accountable. This is important for achieving information trustworthiness. On the other hand, we need to ensure, ideally, that the participant who leaked confidential information should be held accountable. This is important for realizing information compliance management.

Enforcing privacy protection. There are several aspects for privacy protection in the context of secure provenance management. First, how can we protect the privacy of the participants? This is important in, for example, intelligence applications where the source of information may be deemed more important than that information itself. What makes it more challenging is that privacy-protection should be fulfilled without jeopardizing the aforementioned properties, especially accountability. Second, how can we manage the information compliance of confidential information while protecting the privacy (or anonymity) of the honest participants? Third, how can we compute information trustworthiness without jeopardizing the privacy of both data and the participants (including both the source and the intermediate nodes)?

III. A Framework for Secure Provenance Management

The above discussions clearly shows that developing a full-fledged secure provenance management system is challenging. Nevertheless, it also suggests us a characterization of solutions to the problem of secure provenance management. In particular, we believe that a secure provenance management system should be (1) policy-neutral, meaning that it can accommodate the policies for managing the access to existing data and the policies for managing the access to data provenance information, although the later may have not been devised yet; (2) application-neutral, meaning that it should allows one to plug-and-play application-specific modules (e.g., semantic similarity between two documents). Moreover, any solution should cover the whole information lifecycle in provenance-aware systems.

As shown in Figure 3, our framework consists of five layers and eight facets, which are elaborated below.

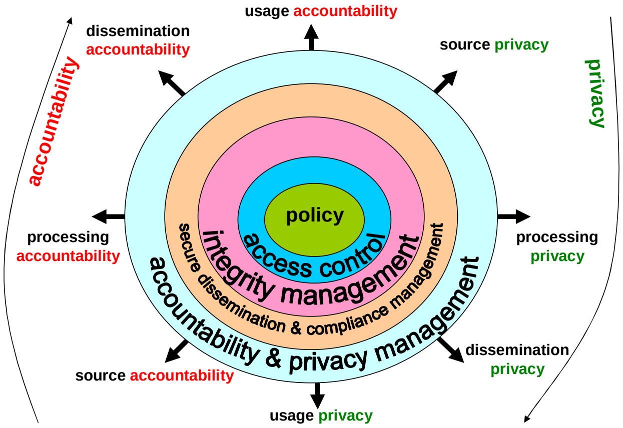

Fig. 3. The framework for secure provenance management systems

Policies. Policies regulate who can operate on which data/information according to what rules. Specifying policies is beyond the scope of the present paper because it is an orthogonal issue.

Access control. Access control should simultaneously accommodate the security needs of data and their provenance information. As discussed in [5], there are three possibilities:

- Data items are as sensitive as their associated provenance information items. In this case, a user is allowed to read the provenance information items if the user is allowed to operates on the corresponding data items. The complication is, however, that the provenance information items is a DAG, and it is not clear whether the user is allowed have access to which portion of the DAG (e.g., only one layer upstream in the DAG?)
- Data items are more sensitive than their associated provenance information items.
- Data items are less sensitive than their associated provenance information items.

In this paper we argue for the need of the following:

- Security techniques, especially access control models, for dealing with the DAG-structured (or even general graphs) provenance data as well as data trustworthiness, dissemination, compliance management. This calls for advanced access control models that can enforce various policies such as authorization policy, delegation policy, trustworthiness policy, dissemination policy, and data usage policy.
- Even for dealing with the DAG-structured data, there could be complications because some data items may be more (less) sensitive than their associated provenance information items. Therefore, it is important to develop a flexible authorization framework able to accommodate the various policies. For example, the provenance information items may themselves be associated with some roles, or mandatary access control security labels. This way, a user may only be authorized to have access to a portion of a DAG.

Integrity management. Integrity of both data and provenance information is important during their storage, processing, and transferring. Ideally, this layer will ensure that integrity is always assured.

Secure dissemination & compliance management. This layer ensures reliable and secure dissemination of data and their associated provenance information. Moreover, it manages the compliance of data and their provenance information.

Accountability & privacy management. This layer enforces accountability of participants and protects the privacy of participants. Since accountability and privacy are at odds with each other, we need solutions that can simultaneously accommodate both accountability and privacy. In particular, both accountability and privacy are relevant in source, processing, dissemination and compliance aspects.

IV. Related Work

The most relevant prior work is Braun et al. [5] who stressed the "DAG-nature" of provenance. In what follows we briefly review other related prior work.

- Prior work on provenance management at individual system level:
  Within the boundary of a single system, provenance management can be addressed at the Operating System (OS) level and the Database Management System (DBMS) level. At the OS level, the Provenance-aware Storage Systems (PASS) project [20] tracks file read/write operations by intercepting system calls. At the DBMS level, the Trio project [1] aims at managing data, data provenance and data uncertainty as one integrated system.
- Prior work on provenance management at the distributed system level:

The Orchestra project [18], [25], [13] has developed techniques that support publishing of dynamical data with provenance-enabled updates. Provenance management in scientific workflow systems has been extensively investigated in [9], [4], [11], [12], [22]. Groth et al. [16] considered an architecture for provenance systems with a service-oriented approach. Dai et al. [10] investigated a method for evaluating data trustworthiness based on the associated provenance in distributed/decentralized systems.

V. Conclusion

We explored the security requirements of secure provenance management systems. As a first step towards tackling the challenges, we presented a characterization of desired solutions to the problem of secure provenance management. As we discussed in the paper, there are many open problems in realizing secure provenance management, ranging from the need of novel access control models to the need of novel accountability and privacy management mechanisms.

Acknowledgement. This work is supported in part by an AFOSR MURI project and by AFOSR grant FA9550-07-1-0041 “Systematic Control and Management of Data Integrity, Quality and Provenance for Command and Control Applications”.

References

[1] P. Agrawal, O. Benjelloun, A. Sarma, C. Hayworth, S. Nabar, T. Sugihara, and J. Widom. Trio: A system for data, uncertainty, and lineage. In *VLDB*, pages 1151–1154, 2006.

[2] O. Benjelloun, A. Sarma, A. Halevy, and J. Widom. Uldbs: Databases with uncertainty and lineage. In *VLDB*, pages 953–964, 2006.

[3] R. Bose and J. Frew. Lineage retrieval for scientific data processing: a survey. *ACM Comput. Surv.*, 37(1):1–28, 2005.

[4] S. Bowers, T. McPhillips, B. Ludäscher, S. Cohen, and S. Davidson. A model for user-oriented data provenance in pipelined scientific workflows. In *International Provenance and Annotation Workshop (IPAW)*, pages 133–147, 2006.

[5] U. Braun, A. Shinnar, and M. Seltzer. Securing provenance. In *HotSec'08*, 2008.

[6] P. Buneman, A. Chapman, and J. Cheney. Provenance management in curated databases. In *SIGMOD'06*, pages 539–550, 2006.

[7] P. Buneman, S. Khanna, and W. Tan. Why and where: A characterization of data provenance. In *Proceedings of the 8th International Conference on Database Theory (ICDT'01)*, pages 316–330, 2001.

[8] J. Cheney. Program slicing and data provenance. *IEEE Data Eng. Bull.*, 30(4):22–28, 2007.

[9] S. Cohen, S. Boulakia, and S. Davidson. Towards a model of provenance and user views in scientific workflows. In *Third International Workshop on Data Integration in the Life Sciences (DILS)*, pages 264–279, 2006.

[10] C. Dai, D. Lin, E. Bertino, and M. Kantarcioglu. An approach to evaluate data trustworthiness based on data provenance. In *5th VLDB Workshop on Secure Data Management*, volume 5159 of *Lecture Notes in Computer Science*, pages 82–98, 2008.

[11] S. Davidson, S. Boulakia, A. Eyal, B. Ludäscher, T. McPhillips, S. Bowers, M. Anand, and J. Freire. Provenance in scientific workflow systems. *IEEE Data Eng. Bull.*, 30(4):44–50, 2007.

[12] J. Golbeck and J. Hendler. A semantic web approach to tracking provenance in scientific workflows. *Concurrency and Computation: Practice and Experience*, 20(5):431–439, 2008.

[13] T. Green, G. Karvounarakis, Z. Ives, and V. Tannen. Update exchange with mappings and provenance. In *VLDB*, 2007.

[14] T. Green, G. Karvounarakis, N. Taylor, O. Biton, Z. Ives, and V. Tannen. Orchestra: facilitating collaborative data sharing. In *SIGMOD'07*, pages 1131–1133, 2007.

[15] P. Groth. *The Origin of Data: Enabling the Determination of Provenance in Multi-institutional Scientific Systems through the Documentation of Processes*. PhD thesis, 2007.

[16] P. Groth, S. Jiang, S. Miles, S. Munroe, V. Tan, S. Tsasakou, and L. Moreau. An architecture for provenance systems. Technical report, Nov. 2006.

[17] R. Hasan, R. Sion, and M. Winslett. Introducing secure provenance: problems and challenges. In *Proceedings of the 2007 ACM Workshop On Storage Security And Survivability (StorageSS)*, pages 13–18, 2007.

[18] Z. Ives, N. Khandelwal, A. Kapur, and M. Cakir. Orchestra: Rapid, collaborative sharing of dynamic data. In *CIDR*, pages 107–118, 2005.

[19] L. Moreau, B. Ludäscher, I. Altintas, R. Barga, S. Bowers, S. Callahan, G. Chin, B. Clifford, S. Cohen, S. Cohen-Boulakia, S. Davidson, E. Deelman, L. Digiampietri, I. Foster, J. Freire, J. Frew, J. Futrelle, T. Gibson, Y. Gil, C. Goble, J. Golbeck, P. Groth, D. Holland, S. Jiang, J. Kim, D. Koop, A. Krenek, T. McPhillips, G. Mehta, S. Miles, D. Metzger, S. Munroe, J. Myers, B. Plale, N. Podhorszki, V. Ratnakar, E. Santos, C. Scheidegger, K. Schuchardt, M. Seltzer, Y. Simmhan, C. Silva, P. Slaughter, E. Stephan, R. Stevens, D. Turi, H. Vo, M. Wilde, J. Zhao, and Y. Zhao. Special issue: The first provenance challenge. *Concurr. Comput. : Pract. Exper.*, 20(5):409–418, 2008.

[20] K. Muniswamy-Reddy, D. Holland, U. Braun, and M. Seltzer. Provenance-aware storage systems. In *Proceedings of the 2006 USENIX Annual Technical Conference*, pages 43–56, 2006.

[21] J. Ockerbloom. Copyright and provenance: Some practical problems. *IEEE Data Eng. Bull.*, 30(4):51–58, 2007.

[22] Y. Simmhan, B. Plale, and D. Gannon. Karma2: Provenance management for data-driven workflows. *Int. J. Web Service Res.*, 5(2):1–22, 2008.

[23] W. Tan. Research problems in data provenance. *IEEE Data Eng. Bull.*, 27(4):45–52, 2004.

[24] W. Tan. Provenance in databases: Past, current, and future. *IEEE Data Eng. Bull.*, 30(4):3–12, 2007.

[25] N. Taylor and Z. Ives. Reconciling while tolerating disagreement in collaborative data sharing. In *SIGMOD'06*, pages 13–24, 2006.

[26] S. Wong, S. Miles, W. Fang, P. Groth, and L. Moreau. Provenance-based validation of e-science experiments. In *International Semantic Web Conference (ISWC)*, pages 801–815, 2005.