Data Security – Challenges and Research Opportunities

Elisa Bertino(✉)

Cyber Center, CS Department, and CERIAS, Purdue University,
West Lafayette, IN, USA
bertino@cs.purdue.edu

Abstract. The proliferation of web-based applications and information systems, and recent trends such as cloud computing and outsourced data management, have increased the exposure of data and made security more difficult. In this paper we briefly discuss open issues, such as data protection from insider threat and how to reconcile security and privacy, and outline research directions.

1 Introduction

Issues around data confidentiality and privacy are under greater focus than ever before as ubiquitous internet access exposes critical corporate data and personal information to new security threats. On one hand, data sharing across different parties and for different purposes is crucial for many applications, including homeland security, medical research, and environmental protection. The availability of "big data" technologies makes it possible to quickly analyze huge data sets and is thus further pushing the massive collection of data. On the other hand, the combination of multiple datasets may allow parties holding these datasets to infer sensitive information. Pervasive data gathering from multiple data sources and devices, such as smart phones and smart power meters, further exacerbates this tension.

Techniques for fine-grained and context-based access control are crucial for achieving data confidentiality and privacy. Depending on the specific use of data, e.g. operational purposes or analytical purposes, data anonymization techniques may also be applied. An important challenge in this context is represented by the *insider threat*, that is, data misuses by individuals who have access to data for carrying on their organizational functions, and thus possess the necessary authorizations to access proprietary or sensitive data. Protection against insider requires not only fine-grained and context-based access control but also anomaly detection systems, able to detect unusual patterns of data access, and data user surveillance systems, able to monitor user actions and habits in cyber space – for example whether a data user is active on social networks. Notice that the adoption of anomaly detection and surveillance systems entails data user privacy issues and therefore a challenge is how to reconcile data protection with data user privacy. It is important to point out that when dealing with data privacy, one has to distinguish between *data subjects*, that is, the users to whom the data is related, and *data users*, that is, the users accessing the data. Privacy

W. Jonker and M. Petković (Eds.): SDM 2013, LNCS 8425, pp. 9–13, 2014.
DOI: 10.1007/978-3-319-06811-4_2,

of both categories of user is important, even though only few approaches have been proposed for data user privacy [6, 8, 9].

Data security is not, however, limited to data confidentiality and privacy. As data is often used for critical decision making, data trustworthiness is a crucial requirement. Data needs to be protected from unauthorized modifications. Its provenance must be available and certified. Data must be accurate, complete and up-do-date. Comprehensive data trustworthiness solutions are difficult to achieve as they need to combine different techniques, such as digital signatures, semantic integrity, data quality techniques, as well taking into account data semantics. Notice also that assuring data trustworthiness may require a tight control on data management processes which has privacy implications.

In what follows we briefly elaborate on the above issues and research challenges.

2 Access Control and Protection from Insider Threat

From a conceptual point of view, an access control mechanism typically includes a reference monitor that checks that requested accesses by *subjects* to protected *objects* to perform certain actions on these objects are allowed according to the access control policies. The decision taken by the access control mechanism is referred to as *access control decision*. Of course, in order to be effective access control mechanisms must support fine-grained access control that refers to finely tuning the permitted accesses along different dimensions, including data object contents, time and location of the access, purpose of the access. By properly restricting the contexts of the possible accesses one can reduce improper data accesses and the opportunities for insiders to steal data. To address such a requirement, extended access control models have been proposed, including time-based access control models, location-based access control models, purpose-based access control models, and attribute-based access control models that restrict data accesses with respect to time periods, locations, purpose of data usage, and user identity attributes [8], respectively.

Even though the area of access control has been widely investigated [2], there are many open research directions, including how to reconcile access control with privacy, and how to design access control models and mechanisms for social networks and mobile devices. Many advanced access control models require that information, such as the location of the user requiring access or user identity attributes [3], be provided to the access control monitor. The acquisition of such information may result in privacy breaches and the use of cloud for managing the data and enforcing access control policies on the data further increases the risks for data users of being target of spear phishing attacks. The challenge is how to perform access control while at the same time maintaining the privacy of the user personal and context information [6, 8].

Social networks and mobile devices acquire a large variety of information about individuals; therefore access control mechanisms are needed to control with which parties this information is shared. Also today user owned mobile devices are increasingly being used for job-related tasks and thus store enterprise confidential data. The main issue is that, unlike conventional enterprise environments in which administrators and other specialized staff are in charge of deploying access control

policies, in social networks and mobile devices end-users are in charge of deploying their own personal access control policies. The main challenge is how to make sure that devices storing enterprise confidential data enforce the enterprise access control policies and to make sure that un-trusted applications are unable to access this data.

It is important to point out that access control alone may not be sufficient to protect data against insider threat as an insider may have a legitimate permission for certain data accesses. It is therefore crucial to be able determine whether an access, even though is granted by the access control mechanism, is "anomalous" with respect to data accesses typical of the job function of the data user and/or the usual data access patterns. For example, consider a user that has the permission to read an entire table in a database and assume that for his/her job function, the user only needs to access a few entries a day and does so during working hours. With respect to such access pattern, an access performed after office hours and resulting in the download of the entire table would certainly be anomalous and needs to be flagged. Initial solutions to anomaly detection for data accesses have been proposed [5]. However these may not be effective against sophisticated attacks and needs to be complemented by techniques such as separation-of-duties [1] and data flow control.

3 Data Trustworthiness

The problem of providing "trustworthy" data to users is an inherently difficult problem which often depends on the application and data semantics as well as on the current context and situation. In many cases, it is crucial to provide users and applications not only with the needed data, but with also an evaluation indicating how much the data can be trusted. Being able to do so is particularly challenging especially when large amounts of data are generated and continuously transmitted. Solutions for improving data, like those found in data quality, may be very expensive and require access to data sources which may have access restrictions, because of data sensitivity. Also even when one adopts methodologies to assure that the data is of good quality, attackers may still be able to inject incorrect data; therefore, it is important to assess the damage resulting from the use of such data, to track and contain the spread of errors, and to recover. The many challenges for assuring data trustworthiness require articulated solutions combining different approaches and techniques including data integrity, data quality, record linkage [4], and data provenance [10]. Initial approaches for sensor networks [7] have been proposed that apply game theory techniques with the goal of determine which sensor nodes need to be "hardened" so to assure that data has a certain level of trustworthiness. However many issues need to be addressed, such as protection again colluding attackers, articulated metrics for "data trustworthiness", privacy-preserving data matching and correlation techniques.

4 Reconciling Data Security and Privacy

As already mentioned, assuring data security requires among other measures creating user activity profiles for anomaly detection, collecting data provenance, and context information such as user location. Much of this information is privacy sensitive and

security breaches or data misuses by administrators may lead to privacy breaches. Also users may not feel comfortable with their personal data, habits and behavior being collected for security purposes. It would thus seem that security and privacy are conflicting requirements. However this is not necessarily true. Notable examples of approaches reconciling data security and privacy include:

- Privacy-preserving attribute-based fine-grained access control for data on a cloud [8]. These techniques allow one to enforce access control policies taking into account identity information about users for data stored in a public cloud without requiring this information to be disclosed to the cloud, thus preserving user privacy.
- Privacy-preserving location-based role-based access control [6]. These techniques allow one to enforce access control based on location, so that users can access certain data only when located in secure locations associated with the protected data. Such techniques do not require however that the user locations be disclosed to the access control systems, thus preserving user location privacy.

Those are just some examples referring to access control. Of course one needs to devise privacy-preserving protocols for other security functions. Recent advances in encryption techniques, such as homomorphic encryption, may allow one to compute functions on encrypted data and thus may be used as a building block for constructing such protocols.

Acknowledgments. The research reported in this paper has been partially supported by NSF under awards CNS-1111512, CNS-1016722, CNS-0964294.

References

1. Bertino, E.: Data Protection from Insider Threats: Synthesis Lectures on Data Management. Morgan & Claypool Publishers, San Rafael (2012)
2. Bertino, E., Ghinita, G., Kamra, A.: Access control for databases: concepts and systems. Found. Trends Databases **3**(1–2), 1–148 (2011)
3. Bertino, E., Takahashi, K.: Identity Management: Concepts, Technologies, and Systems. Artech House, Boston (2010)
4. Inan, A., Kantarcioglu, M., Ghinita, G., Bertino, E.: A hybrid approach to record matching. IEEE Trans. Dependable Sec. Comp. **9**(5), 684–698 (2012)
5. Kamra, A., Bertino, E.: Design and implementation of an intrusion response system for relational databases. IEEE Trans. Knowl. Data Eng. **23**(6), 875–888 (2011)
6. Kirkpatrick, M.S., Ghinita, G., Bertino, E.: Privacy-preserving enforcement of spatially aware RBAC. IEEE Trans. Dependable Sec. Comp. **9**(5), 627–640 (2012)
7. Lim, H.S., Ghinita, G., Bertino, E., Kantarciolgu, M.: A game-theoretic approach for high assurance data. In: Proceedings of the IEEE 28th International Conference on Data Engineering, Washington, DC, USA, 1–5 April 2012
8. Nabeel, M., Shang, N., Bertino, E.: Privacy preserving policy based content sharing in public clouds. IEEE Trans. Knowl. Data Eng. (to appear)

9. Nabeel, M., Shang, N., Bertino, E.: Efficient privacy preserving content based publish subscribe systems. In: Proceedings of the 17th ACM Symposium on Access Control Models and Technologies (SACMAT), Newark, NJ, 20–22 June 2012
10. Sultana, S., Shehab, M., Bertino, E.: Secure provenance transmission for streaming data. IEEE Trans. Knowl. Data Eng. **25**(8), 1890–1903 (2013)