Data Trustworthiness—Approaches and Research Challenges

Elisa Bertino(✉)

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

Abstract. With the increased need of data sharing among multiple organizations, such as government organizations, financial corporations, medical hospitals and academic institutions, it is critical to assess and assure data trustworthiness so that effective decisions can be made based on data. In this paper, we first discuss motivations and relevant techniques for data trustworthiness. We then present an architectural framework for a comprehensive system for trustworthiness assurance and discuss relevant recent work. We highlight open research issues and research directions throughout the paper.

Keywords: Big data · Integrity · Trust management · Security · Policies

1 Introduction

Technology advances and novel software systems, including sensing devices, cyber-physical systems, smart mobile devices, cloud systems, data analytics, and social networks, are making possible to capture, and to quickly process and analyze huge amounts of data from which to extract information critical for society-relevant application domains. Examples of such domains include cyber security, homeland protection, healthcare, energy, transportation, and education. For example, in the security application domain, relevant tasks that can benefit from big data include anomaly detection and user monitoring for protection from insider threat [1]. In homeland protection, by analyzing and integrating data collected on the Internet and Web one can identify connections and relationships among individuals that may in turn help in detecting potential terrorists. By collecting and mining data concerning user travels and disease outbreaks one can predict disease spreading across geographical areas. And those are just a few examples; there are certainly many other application domains where big data can play a major role.

However, in order for analysts and decision makers to produce accurate analysis, make effective decisions and predictions, and take actions data must be trustworthy. Indeed, today's demand for data trustworthiness is stronger than ever. As many organizations are increasing their reliance on data for daily operations and critical decision making, data trustworthiness is arguably one of the most critical issues.

J. Garcia-Alfaro et al. (Eds.): DPM/SETOP/QASA 2014, LNCS 8872, pp. 17–25, 2015.
DOI: 10.1007/978-3-319-17016-9_2

Assuring data trustworthiness is however a difficult problem which often depends on the semantics of the application domain. Solutions for improving data, like those found in data quality, may be very expensive and may require access to data sources which may have access restrictions, because of data sensitivity. Also even when one adopts methodologies to assure that data are of good quality, errors may still be introduced and low quality data be used. Therefore a critical requirement is the ability to assess the trustworthiness of data so to be able to discard untrustworthy data, execute recovery operations to correct data, and strengthen defense measures.

The many challenges of assuring data trustworthiness require articulated solutions combining different approaches and techniques. In this paper we discuss some of those approaches and solutions, and introduce and highlight relevant research challenges. We also describe a cyclic framework for assessing trustworthiness for sensor data streams [2] and extensions to this framework. Throughout the paper we identify and discuss relevant research challenges.

2 Relevant Approaches and Techniques

Currently there is no comprehensive approach to the problem of high assurance data trustworthiness. However, several relevant techniques have been proposed in different areas of the computer science field that can be used as building blocks.

Integrity Models. The Biba integrity model [3] has been the first model specifically designed to assure integrity in information systems. This model is based on a hierarchical lattice of integrity levels, and integrity is defined as a relative measure that is evaluated at the subsystem level. A subsystem is some sets of subjects and data objects. An information system is defined as composed of a number of subsystems. In the Biba model the main integrity threat is that of a subject attempting to improperly change the behavior of another subject by supplying false or incorrect data. Under the Biba model each subject and data object in the system is assigned an integrity level from the hierarchical lattice of integrity levels. An integrity level associated with a subject indicates how much one can trust the subject with respect to supplying trustworthy data. Each data object is also assigned an integrity level, indicating how much the data object can be trusted. Based on such trust levels, the main principle of the Biba model is to prevent a more trusted subject from receiving data supplied by a less trusted subject. This principle then dictates how data access control is enforced. A drawback of the Biba model is that it is not clear how to assign appropriate integrity levels to subjects and data objects and which are the criteria for determining them. An interesting possibility would be to investigate whether reputation techniques [4] could be used to address such issue.

The approach by Clark and Wilson [5] is based on a clear distinction between military security and commercial security. They argue that security policies related to integrity, rather than disclosure, are of the highest priority in commercial information systems and that separated mechanisms are required for the enforcement of these policies. The model by Clark and Wilson has two key notions: well-formed transactions and separation of duty. A well-formed transaction is structured so that a subject cannot manipulate data arbitrarily, but only in constrained ways that ensure internal

consistency of data. Separation of duty requires separating all operations into several subparts and that each subpart be executed by a different subject.

Semantic Integrity. Many commercial DBMS support the specification of conditions, often referred to as *semantic integrity constraints*, which data must satisfy. Examples of such conditions in a demographic database would be that the age of each individual in the database is an integer ranging between 0 and 140, and that the age of an individual must be lower than the ages of his/her living ancestors. Such constraints are used mainly for *data correctness and consistency*. As such semantic integrity techniques are unable to deal with the more complex problem of data trustworthiness in that they are not able to determine whether some data correctly reflect the real world and are provided by some reliable and accurate data source.

Data Quality. Data quality is a major problem in a wide range of information systems, ranging from data warehousing and business intelligence to customer relationship management and supply chain management. Data quality has been investigated from different perspectives, depending also on the precise meaning assigned to the notion of data quality. Data are of high quality "if they are fit for their intended uses in operations, decision making and planning" [6]. Alternatively, the data are deemed of high quality if they correctly represent the real-world construct to which they refer. Several theoretical frameworks have been proposed for understanding data quality. One framework aims at integrating the product perspective (conformance to specifications) and the service perspective (meeting consumers' expectations) [7]. Another framework is based on semiotics to evaluate the quality of the form, meaning and use of the data [8]. One highly theoretical approach analyzes the ontological nature of information systems to define data quality rigorously [9]. In addition to these more theoretical investigation, a considerable amount of research has been devoted to investigating and describing various categories of desirable attributes (or dimensions) of data quality. These categories commonly include accuracy, correctness, currency, completeness and relevance. Nearly 200 such terms have been identified and there is little agreement on their nature (are these concepts, goals or criteria?), their definitions or measures. Tools have also been developed for analyzing and repairing poor quality data, through the use for example of *record linkage techniques* [10].

Even though data quality is a very relevant to the problem of assessing and assuring data trustworthiness, it is not clear whether data quality methodologies scale for big data. Also such methodologies have been mainly designed to deal with data errors "naturally" introduced by mistakes in the applications and/or as result of human errors. As such they are unable to deal with environments in which malicious parties may carry deliberate data deception attacks. In addition, many such methodologies are based on the idea of correcting the data by using the "original data"; however in applications such as sensor-based applications, the original data may have disappeared by the time one realizes that there are errors in the collected data. Addressing such an issue would require a real-time data quality assessment process and the ability to quickly perform data recovery and correction actions.

Reputation Techniques. Reputation systems represent a key technology for securing collaborative applications from misuse by dishonest entities. A reputation system computes reputation scores about the entities in a system, which helps single out those entities that are exhibiting less than desirable behavior. Examples of reputation systems may be found in several application domains; E-commerce websites such as eBay (ebay.com) and Amazon (amazon.com) use reputation systems to discourage fraudulent activities. The EigenTrust [4] reputation system enables peer-to-peer file sharing systems to filter out peers who provide inauthentic content. The web-based community of Advogato.org uses a reputation system [11] for spam filtering. Reputation techniques can be useful in assessing data sources as shown by recent research [12].

3 A Cyclic Framework for Data Trustworthiness

Basic Approach. A cyclic and provenance-aware trust computation framework was proposed by Lim et al. [2] for data streamed from sensor networks. The goal of such framework is to support a continuous process by which: (a) data continuously streamed from a network of sensors are assessed with respect to their trustworthiness; and (b) sensors are continuously assessed based on the data they provide. In essence the goal of the framework is to assign each data item and sensor a *trust score*, that is, a number ranging in the [0,1] interval. By using such score, a user or application can compare inconsistent data and thus decide which data to use and which ones to discard. Low trust scores assigned to sensors may also be early signs of compromised or malfunctioning sensors.

The proposed framework is based on the heuristic that the more trustworthy data a sensor reports, the higher the sensor's trust score is. Moreover, the trustworthiness of a data item depends on the trust scores of the sensors which passed it towards the server node. The sensors through which a data item has been passed in the sensor network represent the provenance of such data item. By taking into account such interdependency relationship (see Fig. 1) between the trustworthiness of data items and sensors, a cyclic trust assessment process is executed in which the trust scores evolve gradually.

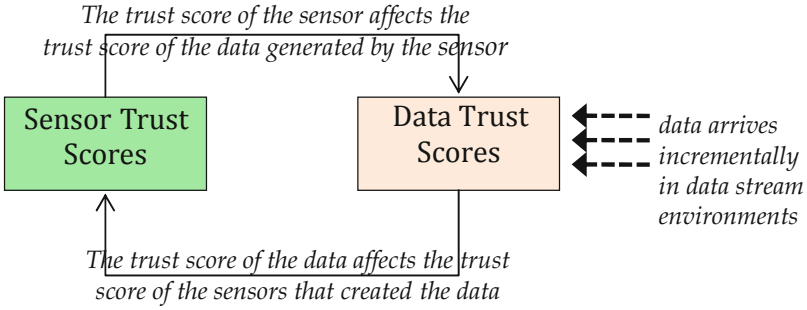

Fig. 1. Interdependency between the data and sensor trust scores

More specifically, in order to reflect the interdependency and continuous evolution properties in computing trust scores, the framework by Lim et al. maintains three different types of trust scores: *current*, *intermediate*, and *next trust scores*. We note that since new data items are continuously added to the stream, executing the cycle once whenever a new data item arrives is enough to reflect the interdependency and continuous evolution properties in the stream environment. The framework works as follows. Trust scores are initially computed based on the values and provenance of data items; we refer to these trust scores as *implicit trust scores*. To obtain these trust scores, two types of similarity functions are used: *value similarity* inferred from data values, and *provenance similarity* inferred from physical provenances. Value similarity is based on the principle that the more data items referring to the same real-world event have similar values, the higher the trust scores of these items are. As most sensor data referring to the same event follow the *normal distribution*, the approach for computing trust scores based on value similarity assumes a normal distribution. A data item that has a value far from the average value computed over all data item values observed during the same time window is thus assigned a lower trust score. Provenance similarity is based on the observation that different physical provenances of similar data values may increase the trustworthiness of data items. In other words, different physical provenances provide more independent data items. For more details on the approach and its experimental evaluation we refer the reader to [2].

A Collusion Attack. As the above framework essentially uses some simple statistical estimators based on value averages, collusions are possible by which several compromised sources collaborate in order to carry out a data deception attack [13]. Such an attack works as follow. Consider a sensor network consisting of 8 sensors all acquiring data about the same environment feature such as humidity. Suppose that a simple statistical test known as 3σ is used, by which values that are higher than three times the standard deviation with respect to the average are discarded. At round 1, all sensors are reliable, and the value accepted by the system (the average among all readings) is close to the actual value (small errors may occur due to device imperfections). At round 2, an adversary compromises three sensors, and alters the readings of these values such that the 3σ interval is skewed towards lower values. Since three distinct sensors report a lower value, the statistical test will conclude that the sensor reporting the highest value must be in error, since its value is outside the confidence interval. Therefore, its value is discarded, and the sensor is marked as less trustworthy for the next round. In the third round, the adversary shifts again its reported values, and manages to make the system to declare the sensor reporting the second highest value untrustworthy as well. This way, through careful selection of reported values, an attacker is able to circumvent the statistical test error detection technique. More importantly, the attacker manages to shift the accepted value far away from the actual value, thus succeeding in the data deception attack.

Protection Against Collusion Attacks. It is important to notice that conventional security approaches like encryption or digital signatures are ineffective against such an attack, as the attacker will alter the data before the data are encrypted and signed by the compromised sensor. Therefore different approaches must be devised.

A promising approach by Rezvani et al. [12, 14] is based on the observation that the above cyclic framework initially assigns the same trust score to all the sensors. Therefore, in order to improve the performance of the cyclic framework [2], Rezvani et al. combine two techniques:

1. *Robust variance estimation for the initial trust score of sensors.* The main idea is to include in the cyclic framework an initial stage in which an initial estimate of two noise parameters for each sensor is obtained; these parameters are bias and variance. Based on such estimate, in the next phase, an initial estimate of the data true values is provided using an estimator inspired by the Maximum Likelihood Estimation (MLE). In the third stage of the proposed framework, the initial estimate of the true values provided in the second stage is used to estimate the trustworthiness of each sensor based on the distance of sensor readings to such initial estimate.
2. *Characterization of the statistical distributions of errors.* It is important to notice that although using the previous technique makes the cyclic framework more robust than its original version which assigns equal trust scores to each sensor, experiments show that the attacker can still skew the results considerably. Thus, the previous technique is extended with a fourth stage based on a novel collusion detection mechanism for eliminating the contributions of the compromised sensors. Such detection mechanism is based on the observation that in a sophisticated collusion attack at least one of the compromised sensors will have highly non stochastic behavior; for example, in the attack scenario by Lim et al. [13], one of the compromised sensors is constrained to reporting values which must be very close to the skewed mean. On the other hand, the error of non-compromised sensors, even when it is large, comes from a large number of independent factors, and thus must roughly have a Gaussian distribution. Consequently, instead of looking just at the Root Mean Square (RMS) magnitude of errors of each sensor, one has to look at the statistical distribution of such errors, assessing the likelihood of whether they came from a normally distributed random variable. Sensors whose errors are highly unlikely to have come from a normally distributed random variable, possibly with a bias, are eliminated. Once the compromised sensors and their readings are eliminated, the noise parameters estimation and the MLE with known variances on the remaining readings are recomputed. Extensive experiments show that this approach is highly effective in detecting colluding attacks. We refer the reader to [12] for details of the approach and its experimental evaluation.

Open Research Issues. In the addition to the problem of collusion, there are many open research issues in the context of the cyclic framework approach which we discuss in what follows.

- *Similarity/dissimilarity of data.* Measuring data similarity is essential in the trust score computation. If we only handle numeric values, the similarity can be easily measured with the difference or Euclidean distance of the values. However, if the value is non-numeric (such as text data), we need to include modeling techniques able to take into account data semantics. For example, if data are names of places, we need to consider spatial relationships in the domain of interest. Possible

approaches that can be used include semantic web techniques, like ontologies and description logics. Similarly, measuring provenance similarity is not easy especially when the provenance is complex. The edit distance which uses the minimum amount of distortion needed to transform one graph into another is the most popular similarity measure in graph theory. However, computing the edit distance for general graphs is known to be an NP-hard problem. Therefore, we need an approximate similarity measure method to efficiently compare graph-shape provenances.

- *Secure and efficient data provenance.* An important requirement for a data provenance trust model is that the provenance information be protected from tampering when flowing across the various parties. In particular, we should be able to determine the specific contribution of each party to the provenance information and the type of modification made (insert/delete/update). We may also have constraints on what the intermediate parties processing the data and providing provenance information can see about provenance information from previous parties along the data provisioning chain. An approach to address such problem is based on approaches for controlled and cooperative updates of XML documents in Byzantine and failure-prone distributed systems [15]. One could develop an XML language for encoding provenance information and use such techniques to secure provenance documents. Also it is critical that provenance be encoded efficiently especially for use in sensor networks. Recent approaches have been proposed based on data dictionary [16] and arithmetic coding techniques [17]. However, they need to be extended to support dynamic wireless networks.
- *Data validation through privacy-preserving record linkage.* In developing solutions for data quality, the use of record linkage techniques is critical. Such techniques allow a party to match, based on similarity functions, its own records with records by another party in order to validate the data. In our context such techniques could be used not only to match the resulting data but also to match the provenance information, which is often a graph structure. Also in our case, we need not only to determine the similarity for the data, but also the dissimilarity of the provenance information. In other words, if two data items are very much similar and their provenance information is very dissimilar, the data item will be assigned a high confidence level. In addition, confidentiality of provenance information is an important requirement because a party may have relevant data but have concerns or restrictions for the data use by another party. Thus application of record linkage technique to our context thus requires addressing the problem of privacy, the extension to graph-structured information, and the development of similarity/dissimilarity functions. Approaches have been proposed for privacy-preserving record linkage [18–20]. However those approaches have still many limitations, such as the lack of support for graph-structured information.
- *Correlation among data sources.* The relationships among the various data sources could be used to create more detailed models for assigning trust to each data source. For example, if we do not have good prior information about the trustworthiness of a particular data source, we may try to use distributed trust computation approaches such as EigenTrust [4] to compute a trust score for the data source based on the trust relationships among data sources. In addition, even if we observe that the same data

is provided by two different sources, if these two sources have a very strong relationship, then it may not be realistic to assume that the data is provided by two independent sources. An approach to address such issue is to develop "source correlation" metrics based on the strength of the relationship among possible data sources. Finally, in some cases, we may need to know "how important is a data sources within our information propagation network" to reason about possible data conflicts. To address such issue one can apply various social network centrality measures such as degree, betweenness, closeness, and information centralities [21] to assign importance values to the various data sources.

4 Conclusions

In this paper we have discussed research directions concerning the problem of providing data that can be trusted by end-users and applications. This is an important problem for which multiple techniques need to be combined in order to achieve good solutions. In addition to approaches and ideas discussed in the paper, many other issues need to be addressed to achieve high-assurance data trustworthiness. In particular, data need to be protected from attacks carried through unsecure platforms, like the operating system, and unsecure applications, and from insider threats. Initial solutions to some of those data security threats are starting to emerge.

Acknowledgments. The work reported in this paper has been partially supported by the Purdue Cyber Center and the National Science Foundation under grant CNS-1111512.

References

1. Bertino, E.: Protection from Insider Threats. Morgan&Claypool, San Rafael (2012)
2. Lim, H.S., Moon, Y.-S., Bertino, E.: Provenance-based trustworthiness assessment in sensor networks. In: Proceedings of the 7th International Workshop on Data Management for Sensor Network (DMSN'10). Singapore (2010)
3. Biba, K.J.: Integrity Considerations for Secure Computer Systems. Technical Report TR-3153, Mitre (1977)
4. Kamvar, S.D., Schlosser, M.T., Garcia-Molina, H.: The eigentrust algorithm for reputation management in P2P networks. In: Twelfth International World Wide Web Conference, pp. 640–651. ACM (2003)
5. Clark, D.D., Wilson, D.R.: A comparison of commercial and military computer security policies. In: Proceedings of IEEE Symposium on Security and Privacy Symposium, Oakland (CA) (1987)
6. Juran, J.M.: Juran on Leadership for Quality—an Executive Handbook. Free Press, New York (1989)
7. Kahn, B., Strong, D., Wang, R.: Information Quality Benchmarks: Product and Service Performance. Communications of the ACM, vol. 45, pp. 184–192. ACM (2002)
8. Price, R., Shanks, G.: A semiotic information quality framework. In: IFIP International Conference on Decision Support Systems: Decision Support in an Uncertain and Complex World. Prato (Italy) (2004)

9. Wand, Y., Wang, R.Y.: Anchoring Data Quality Dimensions in Ontological Foundations. Communications of the ACM, vol. 39, pp. 86–95. ACM (1996)
10. Batini, C., Scannapieco, M.: Data Quality: Concepts, Methodologies and Techniques. Springer (2006)
11. Levien, R:. Attack resistant trust metrics. PhD thesis, University of California—Berkeley, CA, USA (2002)
12. Rezvani, M., Ignjatovic, A., Bertino, E., Jha, S.: Secure data aggregation for wireless sensor networks. IEEE transactions on dependable and secure computing. In press (2014)
13. Lim, H.S., Ghinita, G., Bertino, E., Kantarcioglu, M.: A game-theoretic approach for high-assurance of data trustworthiness in sensor networks. In Proceedings of IEEE 28th International Conference on Data Engineering (ICDE'12), Washington (DC) (2012)
14. Rezvani, M., Ignjatovic, A., Bertino, E., Jha, S.: A robust iterative filtering technique for wireless sensor networks in the presence of malicious attacks. poster abstract. In: Proceedings of ACM Sensys'13 Conference. Rome (Italy) (2013)
15. Mella, G., Ferrari, E., Bertino, E., Koglin, Y.: Controlled and cooperative updates of XML documents in byzantine and failure-prone distributed Systems. ACM Trans. Inf. Syst. Secur. **9**, 421–460 (2006)
16. Wang, C., Hussein, S.R., Bertino, E.: Dictionary based secure provenance compression for wireless sensor networks. IEEE transactions on parallel and distributed systems, in press (2014)
17. Hussein, S.R., Wang, C., Sultana, S., Bertino, E.: Secure data provenance compression using arithmetic coding in wireless sensor networks. In: Proceedings of 33rd IEEE International Performance Computing and Communications Conference (IPCCC 2014), Phoenix (AZ) in press (2014)
18. Scannapieco, M., Figotin, I., Bertino, E., Elmagarmid, A.: Privacy preserving schema and data matching. In: ACM SIGMOD International Conference on Management of Data, pp. 653–664 (2007)
19. Inan, A., Kantarcioglu, M., Bertino, E., Scannapieco, M.: A hybrid approach to private record linkage. In: 24th IEEE International Conference on Data Engineering, pp. 496–505 (2008)
20. Cao, J., Rao, F.-Y., Bertino, E., Kantarcioglu, M.: A hybrid private record linkage scheme: separating differentially private synopses from matching records. In: Proceedings of IEEE 31st International Conference on Data Engineering (ICDE'15), Seoul Korea in press (2015)
21. Jackson, M.O.: Social and Economics Networks. Princeton University Press, Princeton (2008)