Big Data Security and Privacy

Elisa Bertino and Elena Ferrari

Abstract Recent technologies, such as IoT, social networks, cloud computing, and data analytics, make today possible to collect huge amounts of data. However, for data to be used to their full power, data security and privacy are critical. Data security and privacy have been widely investigated over the past thirty years. However, today we face new issues in securing and protecting data, that result in new challenging research directions. Some of those challenges arise from increasing privacy concerns with respect to the use of such huge amount of data, and from the need of reconciling privacy with the use of data. Other challenges arise because the deployments of new data collection and processing devices, such as those used in IoT systems, increase the attack potential. In this paper, we discuss relevant concepts and approaches for Big Data security and privacy, and identify research challenges to be addressed to achieve comprehensive solutions to data security and privacy in the Big Data scenario.

1 Introduction

Technological advances and novel applications, such as sensors, cyber-physical systems, smart mobile devices, cloud systems, data analytics, social networks, Internet of Things (IoT), are making possible to collect, store, and process huge amounts of data, referred to as Big Data, about everything from everywhere and at any time.[1] The Big Data term denotes a data management and analytics paradigm featuring 5V: huge data Volume, high Velocity (i.e., timely response requirements), high Variety of data formats, low Veracity (i.e., uncertainties in the data), and high Value.

[1] Data, data everywhere. The Economist, 25 February 2010, available at http://www.economist.com/node/15557443.

E. Bertino
Computer Science Department, Purdue University, West Lafayette, IN, USA
e-mail: bertino@purdue.edu

E. Ferrari (✉)
Department of Theoretical and Applied Sciences, University of Insubria,
Via Mazzini 5, Varese, Italy
e-mail: elena.ferrari@uninsubria.it

S. Flesca et al. (eds.), *A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years*, Studies in Big Data 31, DOI 10.1007/978-3-319-61893-7_25

Recent advances in sensors, actuators, and embedded computing devices in the physical environment and into physical objects - referred to as Internet of Things (IoT) - further multiply the ability to collect data and act on the physical environment [7]. Gartner forecasts predict that by the year 2020 20.8 billions of IoT devices will be deployed. As IoT grows, so do the volumes of data it generates. CISCO estimates that IoT devices will generate 507.5 zettabytes of data per year by 2019. Moreover, not only today we have technology, such as cloud and high-performance computing systems, for storing and processing huge data sets, we also have advanced data analytics tools that allow one to extract useful knowledge from data and predict trends and events. This will open a number of opportunities for new data-intensive applications in a number of different fields, such as manufacturing and energy management, healthcare management and urban life, just to mention few of them. However, such a scenario increases the threats to the security and privacy of the managed data. Damage and misuse of data affect not only single individuals or organizations, but may have negative impacts on entire social sectors and critical infrastructures. Moreover, smart IoT objects as well as end users are today interconnected by different software platforms. For instance, Online Social Networks (OSNs), represent today the huge example of this trend, with an estimation of around 2.67 billion social media users around the globe by 2018.[2] Such connections multiply the possible threats to security and privacy because they increase the paths on which data may flow. As a result, increasing numbers of attacks have been reported that aim at stealing data through sophisticated attacks, including insider attacks [6].

The problem of data security and privacy is not a new problem; research addressing this problem dates back from the early 70's [23] (see for instance [8] for a short history of research efforts on data security). However, early security and privacy techniques were designed for data stored in corporate database systems and therefore today we need to complement and adapt such early techniques in order to provide full spectrum data protection for Big Data.

In this paper, we first briefly discuss key data security and privacy requirements. We then focus on Big Data and identify key research challenges related to their protection. We then focus on two crucial application domains, namely IoT and OSNs. We finally outline a few concluding remarks. The remainder of the paper is organized as follows. Next section provides an overview of the main data protection requirements. Section 3 illustrates the main research issues in the field of Big Data security and privacy, whereas Sects. 4 and 5 discuss security and privacy issues in the Iot and OSN scenario, respectively. Finally, Sect. 8 concludes the paper.

2 Data Protection Requirements

Traditionally, protecting data requires to ensure three main security properties, that is, data confidentiality, integrity and availability [8], also known as the *CIA triad*. Confidentiality refers to data protection from unauthorized read accesses, whereas

[2] https://www.statista.com.

integrity deals with data protection from unauthorized modifications. Data integrity has been further generalized to data trustworthiness, which refers to making sure not only that data are not modified by unauthorized subjects, but also that data are free from errors, up to date, and originating from reputable sources. Assuring data trustworthiness is thus a difficult problem which often depends on the application domain. Its solution requires combining different techniques, ranging from cryptographic techniques for digitally signing the data, to access control, for checking that only authorized parties modify the data, to data quality techniques, for automatically detecting and fixing data errors [4], provenance techniques [48], for determining from which sources data originate, and reputation techniques, for assessing the reputation of data sources. Finally, availability is the property of assuring that data are available to authorized users. These three requirements are still very critical today, and meeting them is today much more challenging because data attacks are more sophisticated and the data attack surface has expanded, due to increasing data collection activities from many different sources and to data sharing.

In addition to the CIA properties, privacy has emerged as a new critical requirement. Many definitions of data privacy have been so far proposed, and the concept of privacy has evolved over time as a result of the evolution of the means to acquire personal information. One of the first systematic written discussion on the concept of privacy was made by Samuel Warren and Louis Brandeis in their 1890 essay titled "The Right to Privacy" [51], where they define privacy as "the right to be let alone". Warren and Brandeis focused mostly on the press and on the publicity effects produced by the new emerging technological inventions of the time, such as photography and widely distributed newspapers. With the development of more advanced technological products that enabled the acquisition, the carriage, the dissemination, and the persistence of information, such as the video-camera, video-tape, telephone, fax, etc., information privacy continued to attract valuable interest. The appearance and the spread of Internet and of the World Wide Web have made possible to collect massive records of information about individuals (e.g., financial and credit history, medical records, purchase history, telephone calls) that may not exactly know what information is stored about them, by whom, and who has access to it [44]. Today, one of the most used definition of data privacy is due to Allan Westin that defined data privacy as "the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others" [52].

Very often data privacy is seen as the same requirement as data confidentiality, but referring to personal data. There are however relevant differences between the two requirements. It is true that data privacy requires ensuring data confidentiality, because if data are not protected against unauthorized accesses, privacy cannot be ensured. However, privacy has additional issues deriving from the need of taking into account requirements from legal privacy regulations, as well as individual privacy preferences. For example, the concept of purpose is fundamental when dealing with privacy, in that an individual may be fine with sharing his/her own data for research purposes, whereas another individual may not be. Therefore, systems managing privacy-sensitive data may have to collect and record the privacy preferences

concerning the individuals to whom the data refer to. Also data subjects may change their privacy preferences over time. Addressing privacy thus requires, among other things, systems able to enforce not only the access control policies that an organization may have in place to govern accesses to the data, but also data subject preferences and legal regulations. This might require to manage additional dimensions related to the access control decision, such as obligations and user consent. Recently, research has started to address this issue by proposing privacy aware access control systems for relational DBMSs [10, 20, 21] and, more recently for NoSQL databases [17, 18, 35, 37]. However, most of the proposed privacy-enhancing techniques only focus on privacy and do not address the key problem of reconciling data privacy with an effective use of data, especially when the use is for security applications, including cyber security, homeland protection, and health security.

In what follows, we focus on two of the most important data protection requirements, that is, confidentiality and privacy. We then consider two of the most relevant application domains, that is, IoT and social networks.

3 Research Issues in Big Data Confidentiality and Privacy

Several techniques to assure data confidentiality and preserve privacy have been proposed over the last fifteen years, ranging from cryptographic techniques, such as oblivious data structures [50] that hide data access patterns, to data anonymization techniques, that transform data to make more difficult to link specific data records to specific individuals [12], to advanced access control models [19]. However, many research challenges still remain to be addressed. In what follows, we discuss some of them. The presentation in this section is partially based on the discussion in [7].

3.1 Data Confidentiality

Several data confidentiality techniques and mechanisms exist - the most notable being access control and encryption. Both have been widely investigated. However, with respect to access control systems for Big Data we need approaches for:

- **Access control policies merging and integration**. In many cases, Big Data analysis entails integrating data sets originating from multiple, possible heterogeneous, sources; these data sets may be associated with their own access control policies, and these policies must be enforced even when a data set is integrated with other data sets. Therefore, policies need to be integrated and conflicts solved, possibly by using some automated or semi-automated policy integration system [36]. Policy integration and conflict resolution are, however, much more complex when dealing with privacy-aware access control models, as these models allow one to specify policies that include the purpose for which the access to a protected data item

is allowed, obligations arising from the use of data, and special privacy-related conditions that must be meet in order to access the data. Automatically integrating such type of policies and solving conflicts is a major challenge.

- **Authorizations management**. If fine-grained access control is required, manual administration of authorizations on large data sets is not feasible. We need techniques by which authorizations can be automatically granted, possibly based on the user digital identity, profile, and context, and on the data contents and metadata. A first step towards the development of machine learning techniques to support automatic permission assignments to users is by Ni et al. [45]. However, more advanced approaches are needed to deal with dynamically changing contexts and situations.
- **Enforcing access control on Big Data platforms**. Some of the recent Big Data systems allow their users to submit arbitrary jobs encoded in general programming languages. For example, in Hadoop, users can submit arbitrary MapReduce jobs written in Java. This creates significant challenges in order to efficiently enforce fine grained access control for different users. Although there is some initial work [49] that tries to inject access control policies into submitted jobs, more research is needed on how to efficiently enforce such policies in recently developed Big Data stores, especially if access control policies are enforced though the use of fine-grained encryption. Additionally, the variety of data models and query languages adopted by the existing NoSQL datastores make the definition of a general purpose access control mechanism a challenging task. However, some research efforts have been recently started towards the definition of a unifying query language for NoSQL datastores (see e.g., JSONiq [25] and SQL++ [46]) that can be exploited for that purpose. For instance, [17] relies on SQL++ to provide a general approach to support fine grained ABAC (Attribute-based Access Control) within NoSQL platforms. However, more research is needed to define techniques for enforcing fine-grained access control with a reasonable overhead for any query type and policy coverage.

3.2 Privacy

Although solutions to protect data confidentiality represent the core modules for privacy preservation, protecting privacy for Big Data requires to investigate further relevant issues, which include:

- **Techniques to check that data are used for the intended purpose**. The issue here is how to verify that data returned to a user are used for the data owner intended purpose. An initial pioneering approach was proposed in [11] that associates with each data item a set of possible purposes, from an ontology of purposes, for which the data can be used. When a user accesses some data items, the user indicates in the access request the purpose(s) for which the data items are being accessed. The query purposes are then matched against the purposes associated with the data

items to verify that the query purposes comply with the intended use associated with the requested data items. Such an approach needs to be complemented with techniques for automatically and securely identifying the data access purposes, instead of relying on indications given by users as part of their access requests.

- **Support for both personal privacy and population privacy**. In the case of population privacy, it is important to understand what is extracted from the data as this may lead to discrimination [27]. Also when dealing with security with privacy, it is important to understand the tradeoff of personal privacy and collective security.
- **Usability of data privacy policies and user preferences**. Usability is a big issue when dealing with big data privacy and security. Privacy and access control policies must be easily understood by users. We need tools for the average users that help them in specifying their preferences and understand their effect in terms of privacy risks they incur and possible benefits they can get in sharing the data. One direction towards this goal is to empower the user with a secured logical space, a Personal Data Store (PDS) [22] acting as a centralized repository of his/her data. The PDS can then be equipped with a set of analytical tools to reason about the collected data and their sharing with third parties.
- **Privacy-aware access control**. As mentioned before dealing with privacy requires to address further issues wrt data confidentiality, such as obligations, user preferences, and user consent [19]. Although some preliminary work in this direction have been done in the context of Big Data, they mainly focus on specific Big Data platforms (e.g., MongoDB [16, 18]).
- **Risk models**. Different relationships exist between privacy risks and Big Data. One the one hand, Big Data can increase privacy risks, in that they multiply data analysis opportunities; on the other hand, the availability of Big Data sets can reduce security risks in many domains (e.g., national security). The development of models for these two types of risk is critical in order to identify suitable tradeoff and privacy-enhancing techniques to be used.
- **Privacy-aware data lifecycle framework**. A comprehensive approach to privacy for Big Data needs to be based on a systematic data lifecycle approach. Phases in the lifecycle need to be defined and their privacy requirements and implications need to be identified. This is also required by new privacy regulations. For instance, the General Data Protection EU Regulation (GDPR)[3] which has been approved in April 2016 and will enter into application on May 2018 has introduced the privacy by design principle [1]. This will mandatory require that when designing a new system or service that manage personal data, data protection considerations are taken into account starting from the early stages of the design process. Furthermore, the GDPR introduces the Data Protection Impact Assessments (DPIA) which should start prior to the start of processing the personal data, with the goal of identify high risks to the privacy rights of individuals when processing their personal data and possible countermeasures to address them.
- **Data ownership**. The question about who is the owner of a piece of data is often a difficult question. For instance, a notable example is that of photo management

[3] http://europa.eu/rapid/press-release_MEMO-15-6385_en.htm.

in Facebook as users are able to avoid being tagged in a photo,[4] in order to prevent it from being accessible through their profile, but they cannot state how this photo has to be shared in the network. It is perhaps better to replace this concept with the concept of stakeholder. Multiple stakeholders can be associated with each data item. The concept of stakeholder ties well with risks. Each stakeholder would have different (possibly conflicting) objectives and this can be modeled according to multi-objective optimization. In some cases, a stakeholder may not be aware of the others. For example, a user to whom a data item pertains (and thus a stakeholder for the data item) may not be aware that a law enforcement agency is using this data item. Although some preliminary work on this issue has been done in the context of OSNs (see e.g., [30] [31]), technological solutions for Big Data platforms still need to be investigated on support of multiple stakeholder policies.

- **Privacy versus security tradeoff**. The problem of how to reconcile privacy and security is today a major challenge. However, to date very few approaches have been proposed that are suitable for large scale datasets. An example of an initial approach along such direction is the scalable protocol for privacy-preserving data matching by Cao et al. [13] which combines secure multiparty computation (SMC) techniques and differential privacy [24] to address scalability issues.

4 IoT Data Security and Privacy

IoT represents an important emerging trend that, according to various forecasts, will have a major economic impact. IoT applications are changing and improving our every-day lives in a variety of forms, such as with wearable devices that track our sport activities and health status, or with smart home technologies supporting home automation services. However, while on one side, IoT will make many novel applications possible, on the other side, it increases the risks of cyber security attacks to data. In addition, because of its fine-grained, continuous, and pervasive capabilities for data acquisition and control/actuation capabilities, IoT raises concerns about privacy and safety. The OWASP Internet of Things Project[5] has shown that many IoT vulnerabilities arise because of the lack of adoption of well-known security techniques, such as encryption, authentication, access control. This is due to a variety of reasons, such as the cost of deploying privacy and security solutions or the security and privacy unawareness by IT companies involved in the IoT space. But one fundamental reason is because existing security and privacy techniques, tools, and products may not be easily deployed to IoT devices and systems, for reasons such as the variety of hardware platforms and limited computing resources of many types of IoT devices, as well as the underlying decentralized architecture. Therefore, addressing IoT data security and privacy requires extending or re-engineering existing solutions as well as to develop new solutions to fit the specific requirements of

[4]Facebook Help Center - Tag Review. https://www.facebook.com/help/247746261926036/.

[5]https://www.owasp.org/index.php/OWASP_Internet_of_Things_Project.

IoT. Such solutions must ensure protection while data are transmitted and processed at the devices. In addition, in many cases, data availability is critical and therefore solutions minimizing data losses must be devised. In what follows, we survey some projects that cover different aspects of Iot data security and privacy [5].

4.1 Cryptographic Protocols

When dealing with very large IoT systems, one of the key research challenges is to devise efficient and scalable encryption scheme, able to cope with smart objects with very limited processing capabilities. An example of research project in this direction is the certificate-less sign and encryption protocol proposed in [47], that is, a protocol not requiring key certificates and supporting both message encryption and authentication. The protocol works for many different devices, including Raspberry Pi2, and Android. As this protocol does not use expensive pairing operations, it is highly efficient compared to other similar protocols. Other projects along this line have been proposed for specific Iot scenarios, such as for instance the protocol proposed in [43] for efficient authentication operations for networked vehicles. The protocol is able to manage multiple concurrent authentication operations with real-time response time. Response time is critical in that, if a vehicle has to stop suddenly, information about this event has to reach the other vehicles in a very short time so that these vehicles have enough time to break. Therefore, it is crucial that authentication operations both at the sender and the receivers have minimal overhead. To address such requirement, the implementation of the authentication operations takes advantage of the GPU usually present in systems-on-chips today used in vehicles. Another interesting project focuses on encryption protocols for networks consisting of small sensors and drones [53]. In such networks, sensors are on the ground and acquire data of interest from the environment, whereas drones fly over the sensors to collect and aggregate data from them. The main issue here is to save energy and to make sure that drones do not have to wait too long for sensors to start generating encryption keys. To address such requirements, the approach is to use low power listening (LPL) techniques at the sensors and dual radio channels at the drones. In this way, the sensors can timely start generating the cryptographic keys when drones approach.

Results from those projects show that a careful engineering is critical to the effective deployment of cryptographic protocols in IoT. In particular, it is critical to analyze in details the protocols in order to determine the expensive operations so to replace or optimize them, and to understand how to take advantage of specific hardware features of the devices in order to enhance the implementation of the different steps of the protocols.

4.2 Network Security

Security techniques at network level are critical in order to minimize data losses. Such minimization is crucial for many applications, such as monitoring applications and control systems. In order to minimize data losses, it is critical to be able to quickly diagnose the cause of data packet losses so to repair the network as soon as possible. A recent project [40] has addressed this requirement by developing a Fine-Grained Analysis (FGA) tool that investigates packet losses and reports their most likely cause. Such FGA tool is based on profiling the wireless links between the nodes as well as their neighborhood, by leveraging resident parameters, such as RSSI and LQI, available within every received packet. By using those profiles, the tool is able to determine whether the cause of a packet loss is a link that has been jammed or a sensor that has been compromised. In the former case, the FGA tool is able to quite reliably detect the source of interference. The design of the system is fully distributed and event-driven, and its low overhead makes it suitable for resource-constrained entities such as wireless motes. This project is however just an initial approach. Research is needed to develop more advanced FGA tools able to deal with mobile systems and heterogeneous communication technologies which may require using different profiling parameters.

4.3 Application Security

Protecting applications is crucial for data security as attacks to steal data often use application vulnerabilities as stepping stones. It is important to notice that even though today we have several techniques for program analysis and hardening, such techniques need substantial extensions to fit IoT devices. The research in this direction is still in its infancy. A first example of research projects in this area is represented by techniques to protect programs against code injection attacks and code reuse attacks [26]. A well-known approach to protect against those attacks is to instrument the application binary code by inserting a static check statement before any instruction that modifies the program counter. Such check verifies that the target address, to which the program execution has to move, is the correct address, that is, that the next instruction to be executed is the expected one and not an instruction to which the attacker is trying to redirect the execution. Such techniques have been shown to be quite efficient. As the run-time overhead ranges between 0.51 and 12.22%, based on the benchmarked applications [26]. However, the instructions that can modify the program counter are different for different platforms; such variations thus require devising specific instrumentation techniques for specific platforms. Other relevant attacks are those exploiting memory vulnerabilities. An approach for applications written in a variant of the C language specific for TinyOS applications has been proposed in [41]. Such an approach statically analyzes an application to identify memory vulnerabilities. As in some cases it is not possible to statically determine

if a certain piece of code will lead to a vulnerability at run-time, the approach adds some code to check at run-time whether a vulnerability occurs. Also in this project the main issue is to minimize the run-time overhead as this is critical for devices with limited capabilities. Both those projects show that significant work is required to modify existing application program security techniques for use in IoT systems.

4.4 Privacy and Access Control

IoT is more and more evolving into a loosely coupled, decentralized system of cooperating *smart objects*, where high-speed data processing, analytics and shorter response times are becoming more necessary than ever. Such decentralization has a great impact on the way personal and sensitive information generated and consumed by smart objects should be protected, because, without a centralized data management entity, it is more difficult to control how data generated by smart objects are combined and used, even to infer new information. In this scenario, there is the need of defining new enforcement mechanisms for both access control policies and privacy preferences. In this respect, some proposals have recently emerged. For instance, [29] proposed a distributed capability-based access control mechanism exploiting public key cryptography to share information among smart objects. In [39], a two layered architecture is proposed for protecting users' privacy in smart city applications: a first trusted layer, where information is stored and processed by the platform's components, and an open and untrusted second layer, where only generic and unidentifiable information is made available to the external applications. [14] proposed a system for specifying and enforcing privacy preferences in the IoT scenario. The framework provides an expressive language to specify privacy preferences and a mechanism to automatically generate preferences when new information is generated as a result of the data processing. However, the proposal presented in [14] considered a centralized architecture, that is, a scenario where IoT devices have only the capability to sense the data and send them to a data center for being analyzed. In particular, in the framework proposed in [14] sensed data are forwarded by a message broker to a Complex Event Processing system (CEP) as append-only streams of tuples, where registered queries analyze, combine and aggregate them generating new output data. The enforcement monitor statically analyzes every data consumer query and decides if privacy policies of the consumer satisfies the privacy preferences specified by owners of devices generating the data. A challenging issue to be addressed is thus that of designing a fully decentralized privacy enforcement mechanism, where compliance check of data owner privacy preferences is performed directly by smart objects. This has to cope with non negligible overhead that may arise and with the reduced processing capabilities of many smart objects (e.g., sensors). Also in a decentralized setting, the enforcement mechanism should be robust against malicious and colluding smart objects.

5 OSNs Data Security and Privacy

OSNs are one of the most relevant phenomenon in the Big Data area, with billion of users worldwide. OSNs have introduced substantial changes to the way people communicate and socialize within and out of their communities. As a matter of fact, they represent today the biggest available repository of personal information, However, despite all their benefits they also create serious privacy and confidentiality concerns given the nature of information users share over them on almost a daily basis. Users publish their personal stories and updates, and they might also express their opinion by interacting on information shared by others, but, in most cases, they are not fully aware of the size of the audience that gets access to their information. Current commercial OSNs provide very basic form of data protection [15]. In what follows, we survey some of the main security and privacy issues in the realm of OSNs, by covering related relevant projects.

6 Privacy-Aware Access Control

In OSNs, data protection has been mainly approached with Relationship-Based Access Control (ReBAC). According to ReBAC access control decisions are taken by tracking the interpersonal relationships established between users in the network and allowing the formulation of access policies based on them [15]. Privacy settings currently available in commercial OSNs operate under a limited form of ReBAC, but remain both complicated to use, and not flexible enough to model all the privacy preferences that users may require [38]. Most of the mechanisms and techniques that have been suggested for achieving ReBAC in OSNs fall under one of two categories: trust-based or encryption-based. The trust-based approach has mainly been explored under the centralized design of OSNs, where a central entity has full knowledge of the network graph including its nodes, edges, and data ownership, and it is in charge of performing access control. On the other hand, the encryption-based approach has been mostly investigated to address the access control problem under the emerging Decentralized OSNs (DOSNs) scenario. A DOSN is a system that offers OSN services in a peer to peer manner. The concept of DOSNs aims at bringing back control to OSN users and freeing them from the observance of the central service providers. Deploying data encryption to manage access control in DOSNs means that anyone could retrieve the encrypted content but only those who have the corresponding keys can interpret it. This implies that one of the requirements is to offer a mechanism for the distribution, management, and revocation of the corresponding keys, which can introduce a significant overhead due to the huge OSNs population (see e.g., [9, 32]). Whilst such solutions might ensure high data security levels, they have scalability problems and are not flexible enough to support the fine granularity and complex access scenarios required for data dissemination in OSNs. Therefore, what is needed is an investigation of alternative paradigms to perform access control in DOSNs, wrt

the preventive one commonly adopted. A step in this direction is represented by [3], where the authors propose an audit-based mechanism to perform a posteriori access control in DOSNs.

7 Identity Validation

All the access control models and methods discussed thus far assume a mechanism in the system by which subjects have been identified and authenticated. However, identities in OSNs are very loose. To facilitate their adoption and encourage people to join them, only a valid email address is required for a user to create an identity in the OSN. The problem of fake accounts and identity related attacks in OSNs has attracted considerable interest from the research body. An example of research project in this area is SybilLimit [54], that leverages on the fast mixing principle, by which honest nodes should converge to having high connectivity to the rest of the network, to detect Sybil attacks. L. Jin et al. suggest in [34] a framework for the detection of cloning attacks, based on attribute and friends' network similarities. All such approaches, and others following the same approach, aim at detecting malicious nodes that follow identified and formalized attack trends. Another important issue is how to validate identity across multiple social networks. Along this line is the work in [28]. However, despite the many methods so far proposed, real case OSN scenarios demonstrate that malicious activities are still taking a huge share [42]. This is due to the fact that almost all detection mechanisms could catch fake nodes only after they have demonstrated some malicious activity or abnormal behavior [33]. Moreover, this detection tends to fail when fake nodes succeed in establishing enough links with good profiles and imitating normal features and behavior. Therefore, the development of effective methods to detect fake accounts is still an open issue. A complementary promising approach to increase the immunity of OSNs to such threats is to empower their honest users with tools that provide them with guarantees or indications on the trustworthiness of the other peers they want to start interacting with. Along this direction, [2] proposes to exploit the OSN crowd to collaboratively estimate the validity of OSN user identities based only on the information they provide on their profiles.

8 Conclusion

While there is no doubt that the Big Data revolution has created substantial benefits to businesses as well as end users, there are commensurate risks that go along with using Big Data. The need to secure data, to protect private information, being at the same time able to ensure data quality, exists whether data sets are big or small. However, the specific properties of Big Data (volume, variety, velocity, veracity, and value) create new types of risks that necessitate to be addressed. In this paper, we

have highlighted some of them, by also focusing on two key Big Data scenarios, namely Iot and Social Networks. As a final remark, we would like to point out that addressing the today and tomorrow challenges in data security and privacy require multidisciplinary research drawing from many different areas, including computer science and engineering, information systems, statistics, economics, social sciences, political sciences, psychology. We believe that all these perspectives are needed to achieve effective solutions to the problem of security and privacy in the era of Big Data and pervasive data acquisition and use, and especially, to the problem of reconciling security with privacy.

Acknowledgements The work reported in this paper is partially supported by NSF under the grant ACI-1547358.

References

1. T. Antignac, D. Le Metayer, Privacy by design: from technologies to architectures, in *Bart Preneel and Demosthenes* ed. by Ikonomou, Privacy Technologies and Policy, LNCS, vol. 8450 (Springer, Berlin, 2014)
2. L. Bahri, B. Carminati, E. Ferrari, COIP - continuous, operable, impartial, and privacy-aware identity validity estimation for OSN profiles. ACM Trans. Web **10**(4), 23:1–23:41 (2016)
3. L. Bahri, B. Carminati, E. Ferrari, CARDS - collaborative audit and report data sharing for a-posteriori access control in DOSNs, in *Proceedings of the 1st IEEE Conference on Collaboration and Internet Computing (CIC 2015)* (2015)
4. C. Batini, M. Scannapieco, *Data and Information Quality Dimensions, Principles and Techniques* (Springer, Berlin, 2016)
5. E. Bertino, Data privacy for IoT systems: concepts, approaches, and research directions, in *Proceedings of the IEEE International Conference on Big Data (BigData 2016)* (2016)
6. E. Bertino, *Data Protection from Insider Threats*. Synthesis Lectures on Data Management (Morgan & Claypool Publishers, 2012)
7. E. Bertino, Data security and privacy: concepts, approaches, and research directions, in *Proceedings of the 40th IEEE Computer Software and Applications Conference (COMPSAC 2016)* (2016)
8. E. Bertino, R. Sandhu, Database security: concepts, approaches, and challenges. IEEE Trans. Dependable Sec. Comput. **2**(1), 2–19 (2005)
9. O. Bodriagov, G. Kreitz, S. Buchegger, Access control in decentralized online social networks: applying a policy-hiding cryptographic scheme and evaluating its performance, in *Pervasive Computing and Communications Workshops (PERCOM Workshops)* (2014)
10. J.W. Byun, N. Li, Purpose based access control for privacy protection in relational database systems. The VLDB J. **17**(4) (2008)
11. J.W. Byun, E. Bertino, N. Li, Purpose based access control of complex data for privacy protection, in *Proceedings of the 10th ACM Symposium on Access Control Models and Technologies (SACMAT 2005)* (2005)
12. J.W. Byun, A. Kamra, E. Bertino, N. Li, Efficient k-anonymization using clustering techniques, in *Proceedings of the 12th Conference on Database Systems for Advanced Applications (DASFAA 2007)* (2007)
13. J. Cao, E.-Y. Rao, E. Bertino, M. Kantarcioglu, A hybrid private record linkage scheme: separating differentially private synopses from matching records, in *Proceedings of the 31st Conference on Data Engineering (ICDE 2015)* (2015)

14. B. Carminati, P. Colombo, E. Ferrari, G. Sagirlar, Enhancing user control on personal data usage in internet of things ecosystems, in *Proceedings of the IEEE International Conference on Services Computing (SCC 2016)* (2016)
15. B. Carminati, E. Ferrari, M. Viviani, *Security and trust in online social networks*. Synthesis Lectures on Information Security, Privacy, and Trust (Morgan & Claypool Publishers, 2013)
16. P. Colombo, E. Ferrari, Enhancing MongoDB with purpose based access control. IEEE Trans. Dependable Sec. Comput. to appear
17. P. Colombo, E. Ferrari, Towards a unifying attribute based access control approach for NoSQL datastores, in *Proceedings of the 33rd IEEE Conference on Data Engineering (ICDE 2017)*, to appear
18. P. Colombo, E. Ferrari, Towards virtual private NoSQL datastores, in *Proceedings of the 32nd IEEE Conference on Data Engineering (ICDE 2016)* (2016)
19. P. Colombo, E. Ferrari, Privacy aware access control for big data: a research roadmap. Big Data Res. **2**(4), 145–154 (2015)
20. P. Colombo, E. Ferrari, Enforcing obligations within relational database management systems. IEEE Trans. Dependable Sec. Comput. **11**(4), 318–331 (2014)
21. P. Colombo, E. Ferrari, Enforcement of purpose based access control within relational database management systems. IEEE Trans. Knowl. Data Eng. **26**(11), 2703–2716 (2014)
22. Y.A. De Montjoye, E. Shmueli, S.S. Wang, A.S. Pentlan, openPDS: protecting the privacy of metadata through safe answers, in *PLoS One* (2014)
23. D.E. Denning, P.J. Denning, Data security. ACM Comput. Surv. **11**(3), 227–249 (1979)
24. C. Dwork, A. Roth, The algorithmic foundations of differential privacy. Found. Trends Theor. Comput. Sci. **9**(3–4), 211–407 (2014)
25. D. Florescu, G. Fourny, JSONiq: the history of a query language. IEEE Int. Comput. **17**(5) (2013)
26. J. Habibi, A. Panicker, A. Gupta, E. Bertino, DisARM: mitigating buffer overflow attacks on embedded devices, in *Proceedings of the 9th Conference on Network and System Security (NSS 2015)* (2015)
27. S. Hajian, J. Domingo-Ferrer, A. Monreale, D. Pedreschi, F. Giannotti, Discrimination- and privacy-aware patterns. Data Min. Knowl. Discov. **29**(6), 1733–1782 (2015)
28. B.-Z. He, C.-M. Chen, Y.-P. Su, H.-M. Sun, A defense scheme against identity theft attack based on multiple social networks. Expert Syst. Appl. **41**(5), 2345–2352 (2014)
29. J.L. Hernandez-Ramos, D.G. Carrillo, R. Marin-Lopez, A.F. Skarmeta, Dynamic security credentials pana-based provisioning for IoT smart objects, in *Proceedings of the IEEE 2nd World Forum on Internet of Things (WF-IoT'15)* (2015)
30. H. Hu, G.J. Ahn, J. Jorgensen, Multiparty access control for online social networks: model and mechanisms. IEEE Trans. Knowl. Data Eng. **25**, 1614–1627 (2013)
31. P. Ilia, B. Carminati, E. Ferrari, P. Fragopoulou, S. Ioannidis, SAMPAC: socially-aware collaborative multi-party access control, in *Proceedings of the 7th ACM Conference on Data and Applications Security and Privacy (CODASPY 2017)* (2017)
32. S. Jahid, S. Nilizadeh, P. Mittal, N. Borisov, A. Kapadia, Decent: a decentralized architecture for enforcing privacy in online social networks, in *Pervasive Computing and Communications Workshops (PERCOM Workshops)* (2012)
33. J. Jiang, Z.F. Shan, X. Wang, L. Zhang, Y.F. Dai, Understanding sybil groups in the wild. J. Comput. Sci. Technol. **30**(6), 1344–1357 (2015)
34. L. Jin, H. Takabi, J.B. Joshi, Towards active detection of identity clone attacks on online social networks, in *Proceedings of the 1st ACM Conference on Data and Application Security and Privacy* (2011)
35. D. Kulkarni, A fine-grained access control model for key-value systems, in *Proceedings of the 3rd ACM Conference on Data and Application Security and Privacy (CODASPY 2013)* (2013)
36. D. Lin, P. Rao, E. Bertino, N. Li, J. Lobo, EXAM: a comprehensive environment for the analysis of access control policies. Int. J. Inf. Sec. (IJIS) **9**(4), 253–273 (2010)
37. J. Longstaff, J. Noble, Attribute based access control for big data applications by query modification, in *Proceedings of IEEE BigDataService* (2016)

38. M. Madejski, M.L. Johnson, S.M. Bellovin, The failure of online social network privacy settings. Columbia University Academic Commons (2011)
39. O. Mazhelis et al., Towards enabling privacy preserving smart city apps, in *Proceedings of the IEEE Smart Cities Conference* (2016)
40. D. Midi, E. Bertino, Node or Link? Fine-Grained Analysis of Packet Loss Attacks in Wireless Sensor Networks. ACM Trans. Sens. Netw. **12**(2) (2016). Accepted for publication
41. D. Midi, T. Payer, E. Bertino, nesCheck: Memory Safety for Embedded Devices, submitted for publication
42. S. Mitter, C. Wagner, M. Strohmaier, Understanding the impact of socialbot attacks in online social networks. arXiv preprint arXiv:1402.6289 (2014)
43. A.A. Mudgerikar, A. Singla, I. Papapanagiotou, A.A. Yavuz, HAA: hardware-accelerated authentication for internet of things in mission critical vehicular networks, in *Proceedings of the 34th Conference for Military Communications (IEEE MILCOM 2015)* (2015)
44. A. Narayanan, V. Toubiana, S. Barocas, H. Nissenbaum, D. Boneh, A critical look at decentralized personal data architectures. arXiv preprint arXiv:1202.4503 (2012)
45. Q. Ni, J. Lobo, S.B. Calo, P. Rohatgi, E. Bertino, Automating role-based provisioning by learning from examples, in *Proceedings of the 14th ACM Symposium on Access Control Models and Technologies (SACMAT 2009)* (2009)
46. K.W. Ong, Y. Papakonstantinou, R. Vernoux, The SQL++ unifying semi-structured query language, and an expressiveness benchmark of SQL-on-Hadoop, NoSQL and NewSQL databases. CoRR, arXiv:1405.3631 (2014)
47. S.H. Seo, J. Won, E. Bertino, pCLSC-TKEM: a Pairing-free Certificateless Signcryption-tag key encapsulation mechanism for a privacy-preserving IoT. Trans. Data Priv. (2016)
48. S. Sultana, E. Bertino, A Distributed system for the management of fine-grained provenance. J. Database Manag. **26**(2), 32–47 (2015)
49. H. Ulusoy, P. Colombo, E. Ferrari, M. Kantarcioglu, E. Pattuk, GuardMR: fine-grained security policy enforcement for MapReduce systems, in *Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security (ASIACCS'15)* (2015)
50. H.X. Wang, K. Nayak, C. Liu, E. Shi, E. Stefanov, Y. Huang, Oblivious data structures. IACR Cryptology ePrint Archive (2014)
51. S.D. Warren, L.D. Brandeis, The Right to Privacy. Harvard Law Review (1890), pp. 193–220
52. A. Westin, *Privacy And Freedom* (Atheneum, New York, 1967), p. 7
53. J. Won, S.H. Seo, E. Bertino, A secure communication protocol for drones and smart objects, in *Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security (ASIACCS '15)* (2015)
54. H. Yu, P.B. Gibbons, M. Kaminsky, F. Xiao, Sybillimit: a near-optimal social network defense against sybil attacks, in *Proceedings of the IEEE Symposium on Security and Privacy* (2008)