Towards Risk-aware Policy based Framework for Big Data Security and Privacy
(Position Paper)

James Joshi, Balaji Palanisamy
School of Information Sciences, University of Pittsburgh, PA, USA

Abstract: The volume, velocity, variety and the veracity aspects of Big Data introduce significant security and privacy challenges. While the existing approaches to privacy and security, in particular the issue of access and trust management, themselves face significant challenges to managing the immense volumes and heterogeneity of information, the explosive growth of dynamically generated data in various media forms and formats, which may be inaccurate and incomplete, significantly exacerbates these challenges. Further, considering the increasing heterogeneity of data sources that may or may not be reliable and the issue of multiple security domains with varying security and privacy policies/mechanisms, the Big Data essentially presents seemingly insurmountable challenges. It is critical that security and privacy risks through inference, undesirable information flows, inaccurate and incomplete data, etc., need to be properly understood to synthesize appropriate solutions for Big Data. In this position paper, we discuss challenges and motivate towards a need for a risk-aware policy framework to provide a holistic approach to address these challenges. Realizing such a framework involves solving renewed challenges related to policy engineering and risk management approaches and integrating them.

1. Introduction

With the rapid advances in sensor technologies, mobile, cloud and high speed computing, we have already witnessed the emergence of the Big Data era [5]. While the immense volume of data being produced at exponential rates by plethora of increasingly heterogeneous computing devices and sensors, that constitutes the Internet of Things (IoT), show huge potential for the better understanding of various phenomena and events [5, 4, 15] through predictive data analytics, the same data can be significantly misused and exploited for harvesting privacy sensitive information, or affecting the veracity or integrity of results of data analytics that will be used by people and enterprises for decision making purposes. The diversity of data sources introduces immense variety in the media types such as text, images, videos and variations within these; further, establishing reliability and trustworthiness, as well as completeness of data from different sources become very difficult [5]. These exacerbate the problem of ensuring overall quality of data and information related to an individual or an enterprise throughout its lifecycle. The richness of multimedia data generates unique privacy risks as they can be correlated to reveal very sensitive information [16]. The velocity with which multimedia data flows through the networks and devices enabled by rapid advances in mobile and cloud computing, and networking technologies, adds another level of challenge with regards to securely processing potentially inaccurate, unfiltered data in motion [5]. The increasing volume, velocity, and variety of data and the increasing challenge with regards to establishing veracity of such data, thus, present an unprecedented level of security and privacy challenges. In particular, the threat landscape has seen an immense growth resulting in a significant increase in number of threats witnessed in short periods of time. This rapid growth in the threat spectrum have also resulted in many sophisticated hacker tools and cyber criminals now have computing power that never existed before. Seen along with such increase in threats are also many sophisticated hacker tools, which if coupled with emerging big data analytics tools, will enable cyber criminals to acquire computing resources to create large scale security incidents that never existed before.

Critical to devising a holistic solution for security and privacy challenges faced by individuals as well as enterprises is the understanding of what the various security and privacy risks are and how they can be
appropriately assessed and integrated as a key part of such a solution. Various characteristics of Big Data (the 4Vs) introduce various risks. A crucial privacy risk, for instance, is that of re-identification of individuals and inference of privacy sensitive information about them, even when information may have been de-identified [4, 15]. This may be possible because of multiple overlapping data available through multiple sources employing varying levels of security and privacy controls. Similarly, inaccurate data and models may be collected or created related to people because of potentially faulty data sources (various data capture sensors or computing devices) that generate inaccurate data or maintain low integrity/quality data. Analytic results from such data and associated models may be used for decision making purposes without appropriate validation; this may result in significant risks in individual’s safety and well-being, besides raising various security and privacy risks. Because of the scale at which data may be collected, and the possibility of the high rate of flow of such data to various parties with different motives, as well as the potential to use effective data analytics to derive targeted information, significant risks of high impact data breaches is possible [4, 15]. Similarly, the exponential growth of devices with varying capabilities related to providing protection of information that flows through them create significant issues related to secure interoperation and trust management, as well as trustworthiness of data. At the same time, the Big Data provides tremendous opportunities to understand access behaviors of users so as to assess their trustworthiness and construct risk profiles which can be used to fine tune access control and privacy policies.

2. Risk Aware Access Control and Trust Management Approach

We believe that a risk aware access control and trust management approach is the key to addressing many of the crucial issues related to big data security and privacy challenges. More and more, huge volumes of data related to individuals and enterprises are being generated from a huge variety of sources and stored in many different places in cyberspace (e.g., social media, service providers) – this makes the overall task of tracking and managing restricted accesses to information very difficult. For instance, nowadays, personal information in various forms and details, and protected at varying levels of authorization controls, is easily available over the internet – such as user’s profile and information in multiple social media sites. These can be easily harvested to infer private information about a targeted person. Such information can be used to initiate more elaborate privacy attacks; for instance, in social networks, profile cloning attacks, identity clone attacks and other social media feature based attacks can be launched by accumulating and inferring a user’s personal information and also by exploiting the features of applications harvesting the information [9]. To support an individual user (or an enterprise) in securely managing his personal big data (e.g., social media, health related information, profession related information) distributed over other various external application environments, as well as to securely share such data with other external entities in secure and privacy preserving manner, a user-centric, or enterprise-centric policy framework is crucial. Such a framework should have the following capabilities:

- It should be able to capture various content and context based, and situation aware access control policies over multimedia data. In particular, it is essentially impossible to label each protection object and specify policies on them. The size and dynamic nature of personal multimedia data make it imperative to devise content abstractions and policy templates to specify semantic based policy. Use of big multimedia data analytics can help identify potential security and privacy threats that may arise through, data fusion, aggregation and inference – which can be used to potentially specify policies to control sensitive data leakage. Such a framework should be able to specify policies based and attributes and characteristics of both the subjects as well as the protection objects. Various context parameters – such as local, time and social relationships – are important access control decision factors; work in this area has already started to appear. Similarly, expressive policies that capture purpose, obligations and consent are critical to address privacy protection needs [12, 13].
To capture risk of security and privacy violations, it is important to model historical access behavior and patterns of subjects so as to establish trust values for the users or third party or external applications accessing information. For instance, how obligations have been fulfilled over time can indicate a potential malfeasor (e.g., an insider attacker) [12, 13]. Similarly, a trust based role activation scheme that minimizes the risk exposure such as one proposed in [12] could be extended. In particular, it is important to understand the risk of allowing a permission as well as a group of permissions based on the trustworthiness of the subjects accessing the protected resources. Basis to compute risk of inference by using access control relevant entities (permissions, roles, etc.) is crucial [12, 13].

It is also important to understand the risk posed by the external entities such as a social network application environment (e.g., Facebook) or a cloud provider. For instance, it is typically not clear how much risk is posed by Facebook with regards to the privacy of a user’s information. In this case, Facebook system enforces certain system defined policies, and users can specify their privacy preferences; on the other hand, some controls may be undefined as shown in [1]. While a user has an expectation of privacy protection, it is important to understand whether the Facebook environment and his privacy preferences together achieve the privacy expectation he has. Our recent work in [1] provides a basis for seeing what level of privacy controls can be achieved by users in Facebook like social network – and indicates what level of privacy exposure risks may be possible. Further, the user-centric policy component may be also developed to assess the privacy policies of third applications to understand the risks to the users. Use of such knowledge can help users to tune up their policies so as to ensure lower security and privacy risks. In the context of cloud, understanding the risk to data protection from other users as well as the cloud provider themselves is important to help towards better policy design. Role engineering has gained momentum in role based access control research that can be leveraged to develop better policy engineering approaches.

- The policy framework should be able to interact with external applications to ensure proper policy enforcement. In particular, use of cloud providers or external applications that have their own mechanisms can become very challenging with regards to policy enforcement. For instance, personal data may need to be protected based on a temporal access control policy; however, the cloud provider as such may not provide support for that. Hence, the user-centric policy component either needs to mediate each access or generate policy snapshots at different time instances and update the cloud provider side so that it enforces correctly the policy restrictions using its local mechanism [6]. Policy as a service (PaaS) proposed in [6] aim to address such an issue as an initial step. The virtualization layer for the Hadoop Distributed Filesystem proposed in [3] provides an efficient control point for seamlessly introducing such access control mechanisms as part of Hadoop MapReduce systems in Clouds. Privacy protection of data from cloud providers itself is a huge challenge that have generated interest in employing cryptographic techniques such as attribute based encryption.

- Further, such a framework should also provide a basis for enabling secure interoperability or securely sharing sensitive information between interested parties. Secure sharing or interoperation environments could be loosely coupled or tightly coupled [7, 11]. In a tightly coupled environment one can assume tight integration of policies of multiple domains. On the other hand, in the loosely coupled environment the information sharing requirements are typically transient and ad hoc, and policies are not known to partners– which may necessitate policy and trust negotiation [1]. Significant work in trust management and trust negotiation exists; however, these are not able to handle the scale of big data. Policy integration/combination is a huge challenge because of the semantic heterogeneity issue with regards to information being shared by different domains, as well as the diversity of protection requirements and policies [10, 11, 14]. Several existing work have attempted to provide solutions for policy integration and decentralized trust management approaches [10, 11]. However,
the scale of data and the dynamic nature makes these approaches inadequate [9]. More effective and efficient risk based access and trust management approaches are needed.

Developing such a framework has multiple challenges, including: developing better policy engineering approaches founded on risk management activities, formal policy specification and verification/validation approaches, policy integration and analysis approaches, utilizing big data analytics to provide better contextual and situation awareness for policy updates and supporting policy evolution management, as well as more effective and efficient policy enforcement approaches.

3. Conclusion

Secure Management of Big Data with today’s threat spectrum is a challenging problem. While we have been witnessing an overwhelming growth of data in terms of volume, velocity and variety, interestingly, from a security and privacy standpoint, the threat landscape and security and privacy risks have also seen an unprecedented growth. We believe that significant research effort is needed to build a generic architectural framework towards addressing these security and privacy challenges in a holistic manner. In particular, we believe a risk-based policy framework is crucial towards providing such an integrative solution. Various policy engineering and risk management challenges need to be addressed towards realizing such a framework.

References