Chapter 19
Security and Privacy Issues of Big Data

José Moura
Instituto Universitário de Lisboa, Portugal & Instituto de Telecomunicações, Portugal

Carlos Serrão
Instituto Universitário de Lisboa, Portugal & Information Sciences, Technologies and Architecture Research Center, Portugal

ABSTRACT

This chapter revises the most important aspects in how computing infrastructures should be configured and intelligently managed to fulfill the most notably security aspects required by Big Data applications. One of them is privacy. It is a pertinent aspect to be addressed because users share more and more personal data and content through their devices and computers to social networks and public clouds. So, a secure framework to social networks is a very hot topic research. This last topic is addressed in one of the two sections of the current chapter with case studies. In addition, the traditional mechanisms to support security such as firewalls and demilitarized zones are not suitable to be applied in computing systems to support Big Data. SDN is an emergent management solution that could become a convenient mechanism to implement security in Big Data systems, as we show through a second case study at the end of the chapter. This also discusses current relevant work and identifies open issues.

INTRODUCTION

The Big Data is an emerging area applied to manage datasets whose size is beyond the ability of commonly used software tools to capture, manage, and timely analyze that amount of data. The quantity of data to be analyzed is expected to double every two years (IDC, 2012). All these data are very often unstructured and from various sources such as social media, sensors, scientific applications, surveillance, video and image archives, Internet search indexing, medical records, business transactions and system logs. Big data is gaining more and more attention since the number of devices connected to the so-called "Internet of Things" (IoT) is still increasing to unforeseen levels, producing large amounts of data which

DOI: 10.4018/978-1-5225-8897-9.ch019

needs to be transformed into valuable information. Additionally, it is very popular to buy on-demand additional computing power and storage from public cloud providers to perform intensive data-parallel processing. In this way, security and privacy issues can be potentially boosted by the volume, variety, and wide area deployment of the system infrastructure to support Big Data applications.

As Big Data expands with the help of public clouds, traditional security solutions tailored to private computing infrastructures, confined to a well-defined security perimeter, such as firewalls and demilitarized zones (DMZs) are no more effective. Using Big Data, security functions are required to work over the heterogeneous composition of diverse hardware, operating systems, and network domains. In this puzzle-type computing environment, the abstraction capability of Software-Defined Networking (SDN) seems a very important characteristic that can enable the efficient deployment of Big Data secure services on-top of the heterogeneous infrastructure. SDN introduces abstraction because it separates the control (higher) plane from the underlying system infrastructure being supervised and controlled. Separating a network's control logic from the underlying physical routers and switches that forward traffic allows system administrators to write high-level control programs that specify the behavior of an entire network, in contrast to conventional networks, whereby administrators (if allowed to do it by the device manufacturers) must codify functionality in terms of low-level device configuration. Using SDN, the intelligent management of secure functions can be implemented in a logically centralized controller, simplifying the following aspects: enforcement of security policies; system (re)configuration; and system evolution. The robustness drawback of a centralized SDN solution can be mitigated using a hierarchy of controllers and/or through the usage of redundant controllers at least for the most important system functions to be controlled.

The National Institute of Standards and Technology (NIST) launched very recently a framework with a set of voluntary guidelines to help organizations make their communications and computing operations safer (NIST, 2014). This could be achieved through a systematic verification of the system infrastructure in terms of risk assessment, protection against threats, and capabilities to respond and recover from attacks. Following the last verification principles, Defense Advanced Research Projects Agency (DARPA) is creating a program called Mining and Understanding Software Enclaves (MUSE) to enhance the quality of the US military's software. This program is designed to produce more robust software that can work with big datasets without causing errors or crashing under the sheer volume of information (DARPA, 2014). In addition, security and privacy are becoming very urgent Big Data aspects that need to be tackled (Agrawal, Das, & El Abbadi, 2011). To illustrate this, the social networks have enabled people to share and distribute valuable copyrighted digital contents in a very easy way. Consequently, the copyright infringement behaviors, such as illicit copying, malicious distribution, unauthorized access and usage, and free sharing of copyright-protected digital contents, will become a much more common phenomenon. To mitigate these problems, Big Data should have solid solutions to support author's privacy and author's copyrights (Marques & Serrão, 2013a). Also, users share more and more personal data and user generated content through their mobile devices and computers to social networks and cloud services, loosing data and content control with a serious impact on their own privacy. Finally, one potentially promising approach is to create additional uncertainty for attackers by dynamically changing system properties in what is called a cyber moving target (MT) (Okhravi, Hobson, Bigelow, & Streilein, 2014). They present a summary of several types of MT techniques, consider the advantages and weaknesses of each, and make recommendations for future research in this area.