Chapter 84
Differential Privacy Approach for Big Data Privacy in Healthcare

Marmar Moussa
University of Connecticut, USA

Steven A. Demurjian
University of Connecticut, USA

ABSTRACT

This chapter presents a survey of the most important security and privacy issues related to large-scale data sharing and mining in big data with focus on differential privacy as a promising approach for achieving privacy especially in statistical databases often used in healthcare. A case study is presented utilizing differential privacy in healthcare domain, the chapter analyzes and compares the major differentially private data release strategies and noise mechanisms such as the Laplace and the exponential mechanisms. The background section discusses several security and privacy approaches in big data including authentication and encryption protocols, and privacy preserving techniques such as k-anonymity. Next, the chapter introduces the differential privacy concepts used in the interactive and non-interactive data sharing models and the various noise mechanisms used. An instrumental case study is then presented to examine the effect of applying differential privacy in analytics. The chapter then explores the future trends and finally, provides a conclusion.

INTRODUCTION

Big Data analysis influences most aspects of our modern society, such as mobile services, retail, manufacturing, financial services, medicine and life sciences, as well as physical sciences to name a few (Bertino et al., 2011). Scientific research is being revolutionized by Big Data everyday, for instance in bioinformatics with Next Generation Sequencing increasing the size and number of experimental data sets exponentially. In healthcare, Big Data with transforming patient care towards prevention with substantial home-based and continuous form of monitoring available to patients is definitely personalizing

DOI: 10.4018/978-1-5225-7501-6.ch084

healthcare to the benefit of patients. While the potential benefits of Big Data are real and significant, there remain several considerable technical challenges. However, in this broad range of application areas, data is being collected at an unprecedented scale. The emergence and ever increasing emphasis on the big data era means that more and more information on an individual's health, financials, location, and online activity are continuously being harvested, collected, and processed in the cloud and stored in big data repositories. This results in increased concerns regarding the privacy of these large sets of personal data and the loss of an individual's control over his/her sensitive data (Boyd & Crawford, 2012).

The impact of privacy concerns on a big data application is particularly evident in the healthcare domain which has a long established history in requiring that health information technology must comply with the Health Insurance Portability and Accountability Act (HIPAA) for most importantly release of a patient's medical information as well as security and availability as well. HIPAA must also apply to big-data applications for healthcare. This is strongly tied to a movement towards patient controlled access to their medical information with patients able to define the privacy to determine who can see what information at which times. This is evidenced by work that has emphasized granularity and patient control (Sujansky et al., 2010) and a lifetime electronic health record with complete information available anywhere (Caine, 2013). In healthcare there is a need to distinguish levels of security based on the confidentiality and privacy of the data itself and the way that a patient would seek to make such data available to stakeholders. All of these security and privacy concerns must be addressed within big data applications for healthcare as well as in other domains.

This chapter explores the issues related to the security in general and privacy in specific for big data applications, particularly given that the usage of state-of-the-art analytics has explicitly led to growing privacy concerns. As a result, protecting privacy becomes quite harder as information is processed multiple times and shared among multiple diverse entities in the cloud. One example of this problem involves de-identification and anonymization techniques that have been utilized under the false assumption that they allow organizations to reap the benefits of analytics while preserving individuals' privacy. This relies on the assumption that removing certain personal information from a data set would ensure the identity of the users participating in that data set to remain anonymous. However, this has proved to be a misconception as demonstrated by several re-identification and linkage attacks that different data sources harmfully leak private information when combined and when adversaries are able to use some background knowledge, this will be further discussed in the section "Big Data Security and Privacy Issues".

The first focus of this chapter is to explore the utilization of differential privacy to addresses the aforementioned problems in privacy in order to provide confidence to users that their data is carefully controlled. Differential privacy (DWork, 2006) is defined as the application of noise functions of certain characteristics to datasets or query results so that no specifics of individual records present in the original dataset are revealed, while simultaneously allowing the dataset to provide typical big data analytical insights. This constraint allows the various big data analytics mechanisms to behave almost identically on any two datasets that are sufficiently close but only differ by the applied noise mechanism. A formal differential privacy model (DWork, 2006) defined differential privacy as: "the risk to one's privacy should not substantially increase as a result of participating in a statistical database." Differential privacy has recently received increased attention as a general pipeline for the protection of personal information, especially in the fields of big data analytics. The appeal of differential privacy is that there are usually little or no pre-assumptions about a potential attackers pre-existing background knowledge and offers a solid mathematical formulation of the notion of privacy. In contrast to the aforementioned anonymization techniques, the privacy guarantees of differential privacy are rather strong, but can come