Chapter 54
Big Data and Privacy State of the Art

Amine Rahmani
https://orcid.org/0000-0001-7256-1737
University of Algiers 1 Benyoucef Benkhadda, Algeria

ABSTRACT

The phenomenon of big data (massive data mining) refers to the exponential growth of the volume of data available on the web. This new concept has become widely used in recent years, enabling scalable, efficient, and fast access to data anytime, anywhere, helping the scientific community and companies identify the most subtle behaviors of users. However, big data has its share of the limits of ethical issues and risks that cannot be ignored. Indeed, new risks in terms of privacy are just beginning to be perceived. Sometimes simply annoying, these risks can be really harmful. In the medium term, the issue of privacy could become one of the biggest obstacles to the growth of big data solutions. It is in this context that a great deal of research is under way to enhance security and develop mechanisms for the protection of privacy of users. Although this area is still in its infancy, the list of possibilities continues to grow.

INTRODUCTION

The progress of new information technology and knowledge discovery allowing large scale pervasive surveillance over massive data sets had raised the need to such techniques permitting this technology to get more advanced without compromising privacy of users. Nowadays, privacy has been one of the most enduring issues associated with Big Data rising and digital electronic information spreading. There is a growing concern over information privacy. The world of privacy does not rely only on avoiding observation or hiding personal matters and relationships, but it is more than that, it means the ability of sharing information selectively and not publicly. Every day, oceans of data are being collected over the planet using wireless sensors, and intelligent devices. These data are partly sensitive, some of the data are more sensitive than others. Other parts of the data can be considered as sensitive in some cases while not in other cases. People and researchers may have confused between anonymity and privacy.

DOI: 10.4018/978-1-7998-5351-0.ch054

They may seem the same in context, but in reality they are quite different. One's shopping habits are private information but not anonym while authorship of a political tract are anonym data but not private. However, privacy preserving domain is not limited only on sharing information without revealing sensitive knowledge about it, but even more, it expands to the protection of users' rights, the ability to make intimate personal decisions without government interference is considered to be a privacy right, as is protection from discrimination on the basis of certain personal characteristics. This chapter offers a general overview about privacy preserving concepts and techniques focusing on ones used often in Big Data. The remainder of this chapter is addressed to the presentation and discussion of privacy in context, the debates of cybersecurity and privacy, and the presentation of general introduction of different techniques used to maintain privacy of users.

Context of Privacy

In the light of the advancement of Big Data, information technology had become a general threat to privacy. Privacy concerns had raised in the last few years opening new doors and challenges to scientific especially with the developing of such complex and advanced techniques of data mining. The promise of Big Data Analytics is, at first time, to offer information that can be used for good purposes for both individuals and societies. In fact, taking a closer look, we can realise that data mining had been developed in order to allow moving from information about individuals to generalizations that apply to broad classes. That means that data mining in practice should not pose any privacy risk. The real problem resides in the infrastructure used to support it. The more the data is complete and accurate the better are the results. Having complete, comprehensive and accurate data is what causes the raising of privacy issues. In other words, privacy issues rely on the misuse of existed data and not analytical algorithms. This section is addressed to the discussion of the context and outlines of the word "privacy".

A standard definition of privacy in dictionaries is often "freedom from unauthorized intrusion". However, according to the most of privacy laws, it is a concept that is only applied to "individually identifiable data". Regarding to these two definitions, we can define privacy in two extremities. One is that good data is the one gives us an accurate and complete knowledge without revealing any identifiable information about individuals. Meanwhile, in the other extreme, any improvement in our knowledge about individuals could be considered as an intrusion. Even though, data sets represent general information about group of individuals so there is no escape from the fact that analysing it can improve our knowledge about specific persons which requires to measure both knowledge gained and ability of relating these last to specific individuals.

Identifiability of Data

According to HIPAA, an individually non-identifiable data is a data "that does not identify an individual and with respect to which there is no reasonable basis to believe that the information can be used to identify an individual". That means for regular analytical process, the rate of risk of identification on disclosed data must be quite small in either cases alone or in combination with other information. A good example of that is the one given by Latanya Sweeny in (Sweeny, 2002) where she had presented an example of shared medical data without containing names and addresses in the U.S. These data, once it is combined with publicly available voter registrations that contain other information such as birthdays and genders could easily reveal specific information leading to identify persons and their