Chapter 8
Privacy Preserving Data Mining on Unstructured Data

Trupti Vishwambhar Kenekar  
G. H. Raisoni Institute of Engineering and Technology, India

Ajay R. Dani  
G. H. Raisoni Institute of Engineering and Technology, India

ABSTRACT
As Big Data is group of structured, unstructured and semi-structure data collected from various sources, it is important to mine and provide privacy to individual data. Differential Privacy is one the best measure which provides strong privacy guarantee. The chapter proposed differentially private frequent item set mining using map reduce requires less time for privately mining large dataset. The chapter discussed problem of preserving data privacy, different challenges to preserving data privacy in big data environment, Data privacy techniques and their applications to unstructured data. The analyses of experimental results on structured and unstructured data set are also presented.

INTRODUCTION
It is estimated that on an average around 1.09 billion users accessed popular social networking site “Facebook” every day in the month of March 2016 (Anonymous,2016) As per statistics published on site Internet Live statistics state that search engine “Google” carries out 40000 search queries every second(Anonymous,2016). The number of internet users in the world is estimated to be around 3.17 billion (Statista, 2016). Apart from using social networking sites internet users carry out other activities like electronic fund transfers, online purchases, reservations etc. These activities carried out over internet by different users generate large volume of electronic data every day. In last few years volume of electronic data is growing rapidly because of usage number of different devices like smart phones, video recordings in different offices and locations (e.g. Close Circuit Television (CCTV) recordings), mobile phones with cameras, wireless sensor networks etc. All of these have resulted in serious challenges to data privacy.

DOI: 10.4018/978-1-5225-2486-1.ch008
In order to use “Facebook” user needs to create his profile, which consists of information like name, date of birth etc. (also called as data attributes). These data attributes can be categorized into two categories as follows.

1. **Identifying Attributes (IA):** These are attributes which can identify individuals uniquely or help in their identification. The examples of such attributes are Permanent Account Number (PAN) allotted by Income Tax Department of Government of India, Social Security Number (SSN) in USA, Name and Date of Birth etc.

2. **Sensitive Attributes (SA):** These attributes represent sensitive and private information about individuals like medical history, salary details, web browsing history etc. It is expected that only authorized users should be able to view sensitive and private information of users.

When an individual can be identified from the data (e.g. published frequent item sets) the data privacy is said to be violated. When unauthorized users are able to view sensitive information of individuals (say his medical records or financial status) individual’s privacy is violated. The privacy violation is more serious in case unauthorized user is able to link sensitive and private information to individual’s identity.

In the webpage (Adrienne Felt, 2016) of Computer Science Department of University of Virginia, USA the problem of privacy violation with “Facebook” Platform has been described by researcher Adrienne Felt and it is stated that “Facebook” platform allows any “Facebook” user to add different gadgets to his profile. After adding gadgets user can run external third party applications without leaving “Facebook” site. Any data which can be accessed by the user can be accessed by the owners of these applications. The owner of the application can collect data about user’s friends and network and sensitive information like birthdays and can violate data privacy.

The problem of releasing privately held data without identifying individuals is not a new problem (Sweency, 2002). The main purposes of releasing such information can be making data available to researchers, public (for information and other proposes) or can be purely commercial. However in last decade computing environment has changed significantly posing serious data privacy challenges. The main reasons are as follows.

1. In last decade large number organizations (e.g. social networking sites) having private and sensitive information of individuals has emerged. Many of these organizations are willing to share at least part of this information with other companies for commercial purposes. There are many companies who have commercial interest in mining such data.

2. The user profile on social networking site and his network can be commercially useful to many companies.

3. Increasing web activity and usage of different electronic devices and gadgets generate large volume of electronic data (big data environment).The mining of such massive data sets can provide commercially valuable information to many companies.

4. Increase and availability of cheap computing power has made it possible to apply data mining algorithms on very large data sets (big data mining).

5. In addition to this there is emergence of cloud computing platform (Saurabh K.) has also increases data privacy challenge. The popular service models in cloud computing environment (Saurabh K.) are as follows.
   a. Infrastructure as a service (IaaS).