First Credit Seminar Presentation on
"Privacy and Big Data: Issues and Challenges"

By,
Mr. Brijesh B. Mehta
Admission No.: D14CO002

Supervised By,
Dr. Udai Pratap Rao

Computer Engineering Department
S. V. National Institute of Technology, Surat
m.brijesh@coed.svnit.ac.in

21/11/2014
Overview

1. Introduction
   - Privacy
   - Big Data

2. General Architecture of Big Data Analytics
   - Multi Source Big Data Collecting
   - Intra/Inter Big Data Processing
   - Distributed Big Data Storing

3. Privacy Issues in Big Data
   - Privacy in Big Mobile Data

4. Research Challenges with Privacy and Big Data
   - Privacy and Data Mining
   - Top Ten Big Data Security and Privacy Challenges

5. Existing Privacy Preserving Techniques and their Limitations

6. Conclusion and Future Work

7. References
Definition of Privacy: Privacy is the ability of an individual or group to seclude themselves, or information about themselves, and thereby express themselves selectively[1]

Personal information can be classified in four categories [2]

- **Personally Identifiable Information (PII):** name and address
- **Sensitive Information:** religion, health data
- **Usage Data:** web usage
- **Unique Device Identity:** MAC address, RFID tag
**Definition of Big Data:** Big data is an all-encompassing term for any collection of data sets so large and complex that it become difficult to process using traditional data processing applications [3]

**Big Data Characteristic**[4]
- Volume
- Velocity
- Variety

**Big Data Analytics** [5]
- Capture
- Aggregate
- Process
General Architecture of Big Data Analytics

Fig 1 General Architecture of Data Analytics [4]
Multi Source Big Data Collecting

- **Sources of big data** [5]
  - Public Web and Social Media
  - Mobile Applications
  - Surveys
  - Traditional off-line documents scanned by optical character recognition in to electronic form
  - Sensors and Radio-Frequency Identification (RFID) chips
  - GPS chips

- **Classification of online information** [6]
  - **Born Analog** Created by use of some sensor or camera
  - **Born Digital** Created by use of computer

- **Data Fusion** aggregating multiple sources [6]
Fig 2 Map Reduce Architecture [7]
Fig 3 Word Count using Map Reduce [8]
NoSQL database term stands for open source, distributed, and non-relational database[9]

Most common characteristics of NoSQL databases are[10],[11]:

- Simple and flexible non-relational data models
- Ability to scale horizontally over many commodity clusters
- Provide high availability
- Most of them do not support ACID properties of Relational database, instead they support BASE properties (Basically Available, Soft state, Eventually consistence) [12]. However, couchDB[13] supports ACID properties.
NoSQL Data Models and its members:

- **Key-Value Stores**: Memcached, Redis, BerkeleyDB, Voldemort, Riak
- **Column-Family Stores**: Bigtable, Hadoop HBase, SimpleDB, Cassandra
- **Document Stores**: CouchDB, Couchbase server, MongoDB
- **Graph Databases**: Neo4j
Privacy Issues in Big Data

- Edward Snowden Vs NSA [15]
- Surveillance program has divided data into two part:
  - Content Data
  - Context Data (Meta Data)

Fig 5 White House Survey Results [16]
Mobile location data reveal more information about a user than any other data.

Over-collection of mobile data is one of the major concerns:
- brightest flash light android app [17]

Lee Garber[18] has mentioned that Bit Defender, a Romanian security vendor, has analyzed 836,021 Android applications on Google Play Store and found:
- about 33% of apps could reveal location-related data
- about 5% located and opened photos on user’s phone
- approximately 3% reveal users’ email

Mirco Musolesi[19] has mentioned that in June 2013, Facebook had on average, 819 million monthly active mobile users.
Data Provider
- Limit the access
- Trade privacy for benefit
- provide false data

Data Collector, anonymize the data before sending it to data miner, this technique is called privacy preserving data publishing
- In PPDP, attributes of the table is divided as,
  - *Personal Information Identifier (PII)*: Unique ID, Name, Mobile Number
  - *Quasi-identifier (QID)*: Gender, Age, Zip(Postal) Code
  - *Sensitive Attribute (SA)*: Disease, Salary
  - *Non-sensitive Attribute*
Some of the following anonymization operations[20] may apply on the data:

- Generalization
- Suppression
- Anatomization
- Permutation
- Perturbation

Data Miner

- Privacy preserving association rule mining
- Privacy preserving classification
- Privacy preserving clustering

Decision Maker

- Data provenance
- Web information credibility
Big Data Working Group at Cloud Security Alliance has listed top ten big data security and privacy issues as,

1. Scalable and composable privacy preserving data analytics
2. Cryptographically enforced data centric security
3. Granular access control
4. Secure computations in distributed programming frameworks
5. Security best practices for non-relational data stores
6. Secure data storage and transactional logs
7. Granular audits
8. Data provenance
9. End-point validation and filtering
10. Real time security monitoring
Research Challenges with Privacy and Big Data (Top Ten Big Data Security and Privacy Challenges[22]) II

Fig 6 Big Data Eco-system [22]
## Existing Privacy Preserving Techniques and their Limitations

<table>
<thead>
<tr>
<th>Name of the Techniques</th>
<th>Short Description</th>
<th>Limitations with respect to Big Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Privacy Preserving Data Publishing (Anonymization) (De-identification)</td>
<td>Generalization, Suppression, Anatomization, Permutation, and Perturbation are some of the techniques used in PPDP</td>
<td>Works for structured data only</td>
</tr>
<tr>
<td>Privacy Preserving Data Mining</td>
<td>Privacy Preserving Association Rule Mining, Privacy Preserving Classification, Privacy Preserving Clustering</td>
<td>To find and remove sensitive rule from large amount of data is not feasible</td>
</tr>
<tr>
<td>Differential Privacy</td>
<td>Noisy data is being added into the mined results</td>
<td>Risk of re-identification because same data is collected from many sources</td>
</tr>
<tr>
<td>Privacy Preserving Aggregation (Homomorphic Encryption)</td>
<td>Mining operation performed on encrypted data only</td>
<td>It is function specific so we need to write different functions for different tasks which is not feasible</td>
</tr>
</tbody>
</table>
Some of the existing privacy preserving techniques with little modification can be useful to privacy and big data.

As per big data working group, Homomorphic encryption and differential privacy are some of the promising technologies for preserving privacy in big data.

In future, the detailed analysis of privacy preserving techniques like, homomorphic encryption, differential privacy, and de-identification, for privacy and big data will be presented.


Thank You