OUTLINE

• Introduction
• User Role Based Methodology
  ▫ Data Provider
  ▫ Data Collector
  ▫ Data Miner
  ▫ Decision Maker
• Conclusion
Introduction

• Data mining has attracted more attention, because of the popularity of BIG DATA.
• Data mining:
  ▫ discovering interesting patterns
  ▫ knowledge from large amount of data
To achieve useful knowledge:

- Data preprocessing
- **Data transformation**
- Data mining
- Pattern evaluation & presentation
PPDM

- Privacy Preserving Data Mining
- Consideration
  - Sensitive Raw Data
  - Sensitive Mining Result

- Sensitive information :
  - Non accessible to everyone
USER ROLE BASED METHODOLOGY

- Data Provider
- Data Collector
- Data Miner
- Decision Maker
1. DATA PROVIDER

- Two types:
  - Data to data collector
  - Data to data miner
- Privacy Protection
  - Limit the access
  - Trade privacy for benefit
  - Provide false data
LIMIT THE ACCESS

- **Anti-tracking extensions:**
  - To block trackers from collecting the cookies
  - DNT
  - Disconnect, Do not track me, Ghostery...

- **Advertisement and script blockers**
  - Block ads and kill scripts
  - AdBlock Plus, NoScript...

- **Encryption tool**
  - Private online communication
  - MailCloak, TorChat...
TRADE PRIVACY FOR BENEFITS

• Trade off between loss of privacy and benefits brought by participating in data mining.
  ▫ For better shopping experience, disclosure personal information
  ▫ Age, salary, occupation...
  ▫ Selling the data to data collector
PROVIDE FALSE DATA

• To provide falsy data:
  ▫ Using suckpuppets
    • false online identity for pretending to be another person
  ▫ Using fake identity
    • Network eavesdroppers interfered by clone identity
  ▫ Use mask
    • Create and manage aliases(masks)
    • MaskMe
2. DATA COLLECTOR

- Privacy Preserving Data Publishing (PPDP)
- Basics of PPDP
  - Identifier: Id, mobile number ...
  - Quasi-identifier: age, gender ...
  - Sensitive Attribute: salary, disease ...
  - Non-sensitive Attribute: others
- Anonymization: to provide privacy
  - Not modification on sensitive attribute
  - Changing on quasi-identifier
<table>
<thead>
<tr>
<th>Age</th>
<th>Sex</th>
<th>Zipcode</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Female</td>
<td>12000</td>
<td>HIV</td>
</tr>
<tr>
<td>9</td>
<td>Male</td>
<td>14000</td>
<td>dyspepsia</td>
</tr>
<tr>
<td>6</td>
<td>Male</td>
<td>18000</td>
<td>dyspepsia</td>
</tr>
<tr>
<td>8</td>
<td>Male</td>
<td>19000</td>
<td>bronchitis</td>
</tr>
<tr>
<td>12</td>
<td>Female</td>
<td>21000</td>
<td>HIV</td>
</tr>
<tr>
<td>15</td>
<td>Female</td>
<td>22000</td>
<td>cancer</td>
</tr>
<tr>
<td>17</td>
<td>Female</td>
<td>26000</td>
<td>pneumonia</td>
</tr>
<tr>
<td>19</td>
<td>Male</td>
<td>27000</td>
<td>gastritis</td>
</tr>
<tr>
<td>21</td>
<td>Female</td>
<td>33000</td>
<td>flu</td>
</tr>
<tr>
<td>24</td>
<td>Female</td>
<td>37000</td>
<td>pneumonia</td>
</tr>
</tbody>
</table>

(a)

<table>
<thead>
<tr>
<th>Age</th>
<th>Sex</th>
<th>Zipcode</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>People</td>
<td>1*****</td>
<td>HIV</td>
</tr>
<tr>
<td>1</td>
<td>People</td>
<td>1*****</td>
<td>dyspepsia</td>
</tr>
<tr>
<td>1</td>
<td>People</td>
<td>1*****</td>
<td>dyspepsia</td>
</tr>
<tr>
<td>1</td>
<td>People</td>
<td>1*****</td>
<td>bronchitis</td>
</tr>
<tr>
<td>2</td>
<td>People</td>
<td>2*****</td>
<td>HIV</td>
</tr>
<tr>
<td>2</td>
<td>People</td>
<td>2*****</td>
<td>cancer</td>
</tr>
<tr>
<td>2</td>
<td>People</td>
<td>2*****</td>
<td>pneumonia</td>
</tr>
<tr>
<td>2</td>
<td>People</td>
<td>2*****</td>
<td>gastritis</td>
</tr>
<tr>
<td>3</td>
<td>People</td>
<td>3*****</td>
<td>flu</td>
</tr>
<tr>
<td>3</td>
<td>People</td>
<td>3*****</td>
<td>pneumonia</td>
</tr>
</tbody>
</table>

(b)
A social network is usually modeled as a graph
- Vertex: Entity
- Edge: Relationship
ATTACK MODEL

• Mutual Friend Attack
  ▫ Number of mutual friends of two connected individuals
• Friendship Attack
  ▫ Utilizing the degrees of two vertices connected by an edge
• Degree Attack
  ▫ Not only vertex, but also community identity with is known
PRIVACY-PRESERVING PUBLISHING OF TRAJECTORY DATA

• Trajectory:
  ▫ Mobile
  ▫ LBS (Location Based Service)
  ▫ Recommendation about close restaurants

• To realize the privacy-preserving publication
  ▫ Anonymization techniques
<table>
<thead>
<tr>
<th>id</th>
<th>trajectory</th>
<th>id</th>
<th>trajectory</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_1$</td>
<td>$a_1 \rightarrow b_1 \rightarrow a_2$</td>
<td>$t_1$</td>
<td>$a_1 \rightarrow b_1 \rightarrow a_2$</td>
</tr>
<tr>
<td>$t_2$</td>
<td>$a_1 \rightarrow b_1 \rightarrow a_2 \rightarrow b_3$</td>
<td>$t_2$</td>
<td>$a_1 \rightarrow b_1 \rightarrow a_2$</td>
</tr>
<tr>
<td>$t_3$</td>
<td>$a_1 \rightarrow a_3 \rightarrow b_1$</td>
<td>$t_3$</td>
<td>$a_3 \rightarrow b_1$</td>
</tr>
<tr>
<td>$t_4$</td>
<td>$a_3 \rightarrow b_1$</td>
<td>$t_4$</td>
<td>$a_3 \rightarrow b_1$</td>
</tr>
<tr>
<td>$t_5$</td>
<td>$a_3 \rightarrow b_2$</td>
<td>$t_5$</td>
<td>$a_3 \rightarrow b_2$</td>
</tr>
</tbody>
</table>

(a) (b)
3. DATA MINER

• Aim:
  ▫ To prevent sensitive information from appearing in the mining data

• Data mining:
  ▫ Association Rules
  ▫ Classification
  ▫ Clustering
Association Rule Mining

• Aim:
  ▫ To find interesting associations and correlation relationship

• Steps:
  ▫ Find all frequent patterns
  ▫ Generate strong association rules

• To hide rules:
  ▫ Modify original data to generate Sanitized Data (sensitive rules cannot be mined)
Steps of approach
Classification

• Describing important data classes
• Steps:
  ▫ Learning step : use algorithm to build a classifier
  ▫ Using step : use classifier
• Mostly used:
  ▫ Decision tree
  ▫ Naive Bayesian
  ▫ Support Vector Machine
To be private

- Decision Tree:
  - Secure Multi-party Computation (SMC) on data
  - Shamir’s Secret Algorithm

- Naive Bayesian:
  - Add the noise to the parameters of the classifier

- Support Vector Machine:
  - Transform original data into infinite linear combination series
Clustering

- Multiple groups in high similarities
- In the algorithm,
  - computing $k$ clusters on their own private data set
  - computing the distance
    - each data point
    - each of the $k$ cluster centers.
  - randomized cluster centers are exchanged
  - computing the final clustering result
4. DECISION MAKER

• Goals
  ▫ How to prevent unwanted disclosure of sensitive data
  ▫ How to evaluate the credibility of the received mining result
Data Provenance

- Modification applied on the data
- Data provenance
  - Derivation history of the data
- Two approach
  - Network information to directly seek the provenance
  - Reverse flows of information propagation
Web Information Credibility

- Five criteria to differentiate false and true
  - Authority: unknown
  - Accuracy: no accurate data
  - Objectivity: prejudicial
  - Currency: incomplete, missing, ...
  - Coverage: no links
Conclusion

• A user-role based methodology
• Data provider:
  ▫ Limit access, sell the data, falsify the data
• Data collector:
  ▫ Releasing useful data to miner
  ▫ Preserving from attacks by applying the anonymization techniques
Cont.

- **Data miner:**
  - Keep sensitive info undisclosed
  - Certain mining algorithms
- **Decision maker:**
  - To make correct judgement:
    - Provenance
    - Getting true info from false info
References

THANK YOU 😊

QUESTIONS ???