Privacy and Data Mining: New Developments and Challenges

Stan Matwin
School of Information Technology and Engineering
Université d’Ottawa, Canada
stan@site.uottawa.ca

Plan

• Why privacy??
• Classification of Privacy-preserving Data Mining research (PPDM)
• Examples of current PPDM work
• Challenges
Why privacy and data mining?…

• Like any technology can be used for « good » and « bad » purposes …
• It’s Computer Science that has developed these tools, so…
• A moral obligation to develop solutions that will alleviate [potential] abuses and problems

Privacy

• „fuzzy“, over-general concept
  – legal
  – economic
• Security?
Privacy

• Freedom from being watched ("to be left alone")
• …being able to control who knows what about us, and when [Moor]

Privacy

• A CS « perspective»
  – I am a database
  – Privacy is the ability to control the views
• Threats to privacy due to:
  – The Internet
  – Distributed databases
  – Data mining
• « greased » data
…more precisely

- Privacy preservation: what does that mean?
- Given a table of instances (rows), we cannot associate any instance with a given person
- Naive anonymization…
- …is not sufficient, due to pseudo-identifiers

L. Sweeney published this « attack » in 2001:
- anonymized (de-linked) health records of all 135,000 employees+families of the state of Massachusetts was placed on-line
- Electoral list of Cambridge, MA – bought for $20 (54 805 people)
- 69% records are unique wrt birthdate, ZIP; 87% are unique wrt to bday, ZIP, sex…
- Governor’s health records were identified
- …naive anonymization is not sufficient
Other privacy fiascos

• AOL search engine queries published 2006
• Netflix publicly released a data set containing movie ratings of 500,000 Netflix subscribers between December 1999 and December 2005.
• By matching no more than 8 movie ratings and approximate dates, 96% of subscribers can be uniquely identified.

In statistics

• Statistical Disclosure Control
• A table is published, and the whole table has to be protected
• Risk/quality dilemma
• SDC ignores the use of the table
  – Classification
  – Associations
  – Distributed data
Privacy-preserving Data Mining
PPDM

- Data sharing
- Data publishing
- Cloud
- Two main dimensions:
  - What is being protected: data, results?
  - Data centralized or distributed?

PPDM - dimensions

<table>
<thead>
<tr>
<th>Protecting the data</th>
<th>Data centralized</th>
<th>Data distributed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protecting the results</td>
<td>k-anonymization of results [Gianotti/Pedreschi]</td>
<td>[Jiang, Atzori], [Felty, Matwin]</td>
</tr>
<tr>
<td></td>
<td>• generalization/suppression [Sweeney]</td>
<td>• Horizontal/vertical: SMC-based [Clifton], Homomorphic encryption [Wright], Zhang Matwin</td>
</tr>
</tbody>
</table>
|                     | • randomization [Du]/perturbation [Aggrawal] | }
Privacy Goal: $k$-Anonymity

- **Quasi-identifier (QID):** The set of re-identification attributes.
- **$k$-anonymity:** Each record cannot be distinguished from at least $k-1$ other records in the table wrt QID. [Sween98]

<table>
<thead>
<tr>
<th>Raw patient table</th>
<th>3-anonymous patient table</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Job</strong></td>
<td><strong>Sex</strong></td>
</tr>
<tr>
<td>Engineer</td>
<td>Male</td>
</tr>
<tr>
<td>Engineer</td>
<td>Male</td>
</tr>
<tr>
<td>Lawyer</td>
<td>Male</td>
</tr>
<tr>
<td>Musician</td>
<td>Female</td>
</tr>
<tr>
<td>Musician</td>
<td>Female</td>
</tr>
<tr>
<td>Dancer</td>
<td>Female</td>
</tr>
<tr>
<td>Dancer</td>
<td>Female</td>
</tr>
</tbody>
</table>

Homogeneity Attack on $k$-anonymity

- A data owner wants to release a table to a data mining firm for classification analysis on *Rating*

<table>
<thead>
<tr>
<th>Job</th>
<th>Country</th>
<th>Child</th>
<th>Bankruptcy</th>
<th>Rating</th>
<th># Rees</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cook</td>
<td>US</td>
<td>No</td>
<td>Current</td>
<td>0G/4B</td>
<td>4</td>
</tr>
<tr>
<td>Artist</td>
<td>France</td>
<td>No</td>
<td>Current</td>
<td>1G/3B</td>
<td>4</td>
</tr>
<tr>
<td>Doctor</td>
<td>US</td>
<td>Yes</td>
<td>Never</td>
<td>4G/2B</td>
<td>6</td>
</tr>
<tr>
<td><strong>Trader</strong></td>
<td><strong>UK</strong></td>
<td><strong>No</strong></td>
<td><strong>Discharged</strong></td>
<td><strong>4G/0B</strong></td>
<td><strong>4</strong></td>
</tr>
<tr>
<td>Trader</td>
<td><strong>UK</strong></td>
<td><strong>No</strong></td>
<td><strong>Never</strong></td>
<td><strong>1G/0B</strong></td>
<td><strong>1</strong></td>
</tr>
<tr>
<td>Trader</td>
<td>Canada</td>
<td>No</td>
<td>Never</td>
<td>1G/0B</td>
<td>1</td>
</tr>
<tr>
<td>Clerk</td>
<td>Canada</td>
<td>No</td>
<td>Never</td>
<td>3G/0B</td>
<td>3</td>
</tr>
<tr>
<td>Clerk</td>
<td>Canada</td>
<td>No</td>
<td>Discharged</td>
<td>1G/0B</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>24</strong></td>
</tr>
</tbody>
</table>

- Inference: {Trader,UK} $\rightarrow$ fired
- Confidence = $4/5 = 80\%$
- An inference is **sensitive** if its confidence > threshold.
p-Sensitive k-Anonymity

- for each equivalence class EC there is at least \( p \) distinct values for each sensitive attribute
- **Similarity attack** occurs when the values of sensitive attribute

<table>
<thead>
<tr>
<th>Age</th>
<th>Country</th>
<th>Zip Code</th>
<th>Health Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;30</td>
<td>America</td>
<td>142**</td>
<td>HIV</td>
</tr>
<tr>
<td>&lt;30</td>
<td>America</td>
<td>142**</td>
<td>HIV</td>
</tr>
<tr>
<td>&lt;30</td>
<td>America</td>
<td>142**</td>
<td>Cancer</td>
</tr>
<tr>
<td>&lt;30</td>
<td>America</td>
<td>142**</td>
<td>Cancer</td>
</tr>
<tr>
<td>&gt;40</td>
<td>Asia</td>
<td>130**</td>
<td>Hepatitis</td>
</tr>
<tr>
<td>&gt;40</td>
<td>Asia</td>
<td>130**</td>
<td>Pneumonia</td>
</tr>
<tr>
<td>&gt;40</td>
<td>Asia</td>
<td>130**</td>
<td>Asthma</td>
</tr>
<tr>
<td>&gt;40</td>
<td>Asia</td>
<td>130**</td>
<td>Heart Disease</td>
</tr>
<tr>
<td>3+</td>
<td>America</td>
<td>142**</td>
<td>Flu</td>
</tr>
<tr>
<td>3+</td>
<td>America</td>
<td>142**</td>
<td>Flu</td>
</tr>
<tr>
<td>3+</td>
<td>America</td>
<td>142**</td>
<td>Flu</td>
</tr>
<tr>
<td>3+</td>
<td>America</td>
<td>142**</td>
<td>Indigestion</td>
</tr>
</tbody>
</table>

2-Sensitive 4-Anonymity

I-Diversity

- every equivalence class in this table has at least \( l \) well represented values for the sensitive attribute
- **Distinct l-diversity**: the number of distinct values for a sensitive attribute in each equivalence class to be at least \( l \)
- \( l \)-Diversity may be difficult and unnecessary to achieve and it may cause a huge information loss.

<table>
<thead>
<tr>
<th>Zip Code</th>
<th>Age</th>
<th>Nationality</th>
<th>Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1305*</td>
<td>≤ 40</td>
<td>Heart Disease</td>
</tr>
<tr>
<td>4</td>
<td>1305*</td>
<td>≤ 40</td>
<td>Viral Infection</td>
</tr>
<tr>
<td>9</td>
<td>1305*</td>
<td>≤ 40</td>
<td>Cancer</td>
</tr>
<tr>
<td>10</td>
<td>1305*</td>
<td>≤ 40</td>
<td>Cancer</td>
</tr>
<tr>
<td>5</td>
<td>1485*</td>
<td>&gt; 40</td>
<td>Cancer</td>
</tr>
<tr>
<td>6</td>
<td>1485*</td>
<td>&gt; 40</td>
<td>Heart Disease</td>
</tr>
<tr>
<td>7</td>
<td>1485*</td>
<td>&gt; 40</td>
<td>Viral Infection</td>
</tr>
<tr>
<td>8</td>
<td>1485*</td>
<td>&gt; 40</td>
<td>Viral Infection</td>
</tr>
<tr>
<td>2</td>
<td>1306*</td>
<td>≤ 40</td>
<td>Heart Disease</td>
</tr>
<tr>
<td>3</td>
<td>1306*</td>
<td>≤ 40</td>
<td>Viral Infection</td>
</tr>
<tr>
<td>11</td>
<td>1306*</td>
<td>≤ 40</td>
<td>Cancer</td>
</tr>
<tr>
<td>12</td>
<td>1306*</td>
<td>≤ 40</td>
<td>Cancer</td>
</tr>
</tbody>
</table>
t-closeness

- An equivalence class EC is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. [5].

- It solves the attribute disclosure problems of l-diversity, i.e. skewness attack and similarity attack, [6]

<table>
<thead>
<tr>
<th>ZIP Code</th>
<th>Age</th>
<th>Salary</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4767*</td>
<td>≤ 40</td>
<td>3K</td>
</tr>
<tr>
<td>3</td>
<td>4767*</td>
<td>≤ 40</td>
<td>5K</td>
</tr>
<tr>
<td>8</td>
<td>4767*</td>
<td>≤ 40</td>
<td>9K</td>
</tr>
<tr>
<td>4</td>
<td>4790*</td>
<td>≥ 40</td>
<td>6K</td>
</tr>
<tr>
<td>5</td>
<td>4790*</td>
<td>≥ 40</td>
<td>11K</td>
</tr>
<tr>
<td>6</td>
<td>4790*</td>
<td>≥ 40</td>
<td>8K</td>
</tr>
<tr>
<td>2</td>
<td>4760*</td>
<td>≤ 40</td>
<td>4K</td>
</tr>
<tr>
<td>7</td>
<td>4760*</td>
<td>≤ 40</td>
<td>7K</td>
</tr>
<tr>
<td>9</td>
<td>4760*</td>
<td>≤ 40</td>
<td>10K</td>
</tr>
</tbody>
</table>

0.167-closeness w.r.t. salary and 0.278-closeness w.r.t. Disease[5]

Two basic approaches

- Camouflage
- Hiding in the crowd

Data modification/perturbation

K-anonymization
Randomization

- Alice’s age
- Add random number to Age
- 30 becomes 65 (30+35)

Reconstruction (linking)

- initial (confidential) values $x_1, x_2, ..., x_n$ have an (unknown) distribution $X$
- For protection, we perturb them with values $y_1, y_2, ..., y_n$ with a known distribution $Y$
- given
  - $x_1+y_1$, $x_2+y_2$, ..., $x_n+y_n$
  - distribution $Y$
  
Find an estimation of the distribution $X$. 
Works well

Privacy measure

If in the perturbed data, we can identify an original value $x$ in an interval $[x_1, x_2]$ with probability $c\%$, we have a $c\%$ confidence in the privacy of $x$.

<table>
<thead>
<tr>
<th></th>
<th>50%</th>
<th>95%</th>
<th>99.9%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discretization</td>
<td>0.5 x W</td>
<td>0.95 x W</td>
<td>0.999 x W</td>
</tr>
<tr>
<td>Uniform</td>
<td>0.5 x $2\alpha$</td>
<td>0.95 x $2\alpha$</td>
<td>0.999 x $2\alpha$</td>
</tr>
<tr>
<td>Gaussian</td>
<td>$1.34 x \sigma$</td>
<td>$3.92 x \sigma$</td>
<td>$6.8 x \sigma$</td>
</tr>
</tbody>
</table>

example
- Salary 20K - 150K
- 95% confidence
- 50% privacy for uniform distr.
- $2\alpha = 0.5*130K / 0.95 = 68K$

- For a high level of confidence, discretization hurts the results
- Gaussian distribution is better for higher confidence levels
privacy measures

• For modification methods
• First – wrt the interval to which we generalize a value
• We inject ”noise” with a random variable $A$ with distribution $f$
• The privacy measure is

$$\Pi(A) = 2 \int_{\Omega_A} f_A(a) \log_2 f_A(a) da$$

• We measure entropy

Differential privacy

• The desideratum: “access to a database should not enable one to learn anything about individual that could not be learned without access” [Dalenius 77]: similar to semantic security of Goldwasser & Micali
• Impossible because of auxiliary knowledge (AK): database of avg height of people of different nationalities + AK = SM is 2 cm shorter than avg Israeli male
Differential privacy cont’d

• A randomized function $K$ gives $\varepsilon$-differential privacy if for all data sets $D_1$ and $D_2$ differing on at most one element, and all $S \subseteq \text{Range}(K)$,
  
  $\Pr[K(D_1) \in S] \leq \exp(\varepsilon) \times \Pr[K(D_2) \in S]$

• A relative guarantee of non-disclosure: any disclosure is as likely whether or not the individual participates in $D$

• $K$ is a protection ("sanitization") scheme, $\in S$ represents a query about a database

---

Differential privacy cont’d

• For every pair of inputs that differ in one value

• For every output

• Adversary should not be able to distinguish between any $D_1$ and $D_2$ based on any $O$:

$$\log \left( \frac{\Pr(D_1 \rightarrow O)}{\Pr(D_2 \rightarrow O)} \right) < \varepsilon (\varepsilon > 1)$$
Distributed data

- Vehicle/accident data
- To discover the causes of accidents we need to know the attributes of different components from different manufacturers (brakes, tires)
- They will not disclose these values in the open
- Vertical partition

Distributed data

- A medical study carried out in several hospitals
- Would like to *merge* the data for bigger impact of results (results on 20,000 patients instead of 5,000 each)
- For legal reasons, cannot just share then open data
- Horizontal partition
Association Rule Mining Algorithm [Agrawal et al. 1993]

1. \( L_1 \) = large 1-itemsets
2. for \((k = 2; L_{k-1} \neq \emptyset; k + +)\) do begin
3. \( C_k = \text{apriori-gen}(L_{k-1}) \)
4. for all candidates \( c \in C_k \) do begin
5. compute \( c.count \)
6. end
7. \( L_k = \{c \in C_k \mid c.count \geq \text{min-sup}\} \)
8. end
9. Return \( L = \bigcup_k L_k \)

\( c.count \) is the frequency of an itemset.

to compute frequency, we need access to values of attributes belonging to different parties

Example

- \( c.count \) is the scalar product.
- \( A = \) Alice’s attribute vector, \( B = \) Bob’
- \( AB \) is a candidate frequent itemset
- \( c.count = A \cdot B = 3. \)
- How to perform the scalar product preserving the privacy of Alice and Bob?
**Homomorphic Encryption**  
[Paillier 1999]

- Privacy-preserving protocol based on the concept of homomorphic encryption
- The homomorphic encryption property is

\[
e(m_1) \times e(m_2) \times \cdots \times e(m_n) = e(m_1 + m_2 + \cdots + m_n)
\]

- \(e\) is an encryption function \(e(m_i) \neq 0\)
Bob

\[ W_1 = e(A_1 + R_1 \times X) \times B_1 \]
\[ W_2 = e(A_2 + R_2 \times X) \times B_2 \]
\[ \cdots \]
\[ W_N = e(A_N + R_N \times X) \times B_N \]

\[ B_i = 0 \Rightarrow W_i = 0 \]
\[ B_i = 1 \Rightarrow W_i = e(A_i + R_i \times X) \times B_i = e(A_i + R_i \times X) \]

Bob computes
\[ W' = \prod_{j=0}^{N} [W_j \mod X] \]
\[ = \prod_{j=0}^{N} [e(A_j + R_j \times X) \mod X = [e(A_{N_j} + \cdots + A_j + (R_{N_j} + \cdots + R_j + R') \times X] \mod X \]
encrypts, sends to Alice.

**Last stage**

- Alice decrypts \( W' \) and computes modulo \( X \).

\[ c.\text{count} = d(e(A_1 + A_2 + \cdots + A_j + (R_1 + R_2 + \cdots + R_j + R') \times X)) \mod X \]

\( (A_1 + A_2 + \cdots + A_j) \leq N < X \)

\( ((R_1 + R_2 + \cdots + R_j + R') \times X) \mod X = 0 \)

- She obtains \( A_1 + A_2 + \cdots + A_j \) for these \( A_j \) whose corresponding \( B_j \) are not 0, which is \( = c.\text{count} \)
- Privacy analysis
Now looking at data mining results...

Can data mining results reveal personal information? In some cases, yes: [Atzori et al. 05]:

An association rule:

$$a_1 \land a_2 \land a_3 \Rightarrow a_4 \left[ \text{sup} = 80, \text{conf} = 98.7\% \right]$$

Means that

$$\sup(\{a_1, a_2, a_3, a_4\}) = 80$$

So

$$\frac{\sup(\{a_1, a_2, a_3\})}{0.987} = \frac{0.8}{0.987} = 81.05$$

And $a_1 \land a_2 \land a_3 \land \lnot a_4$ has support =1, and identifies a person!!

Protecting data mining results

- A $k$-anonymous patterns approach and an algorithm (inference channels) detect violations of $k$-anonymity of results
Discrimination and data mining

- [Pedreschi et al 07] shows how DM results can lead to discriminatory rules
- In fact, DM’s goal is discrimination (between different sub-groups of data)
- They propose a measure of potential discrimination with lift: to what extent a sensitive is more assigned by a rule to a sensitive group than to an average group

Other challenges

- Privacy and social networks
- Privacy definition – where to look for inspiration (economics?)
- Text data – perturbation/anonymization methods don’t work
- Medical data: trails [Malin], privacy of longitudinal data
- Mobile data -
GeoPKDD

• European project on Geographic Privacy-aware Knowledge Discovery and Delivery
• Data from GSM/UMTS and GPS
First obtaining spatio-temporal trajectories, then patterns

Trajectory = sequence of points visited in a temporal order

Pattern = set of frequent trajectories with similar transition times

Privacy of spatio-temporal data

- Modify the data in such a way each trajectory be indistinguishable from k other trajectories
- … by minimizing distortion introduced into the data
Conclusion

• A major challenge for database/data mining research
• Lots of interesting contributions/papers, but lack of a systematic framework
• …?