Maintaining Data Privacy in Association Rule Mining

Shariq Rizvi
Indian Institute of Technology, Bombay

Joint work with:

Jayant Haritsa
Indian Institute of Science
A Typical Web-Service Form
The Good Side

• Better aggregate models

  “Action movies released in *July* rarely bomb at the box office”

• Improved customer services

  “amazon.com: If you are buying *Macbeth*, you may want to read *The Count of Monte Cristo*”
The Dark Side

- Breach of data privacy

<table>
<thead>
<tr>
<th>Major Illnesses</th>
<th>YES</th>
<th>NO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Myopia</td>
<td></td>
<td>V</td>
</tr>
<tr>
<td>Lung Cancer</td>
<td>V</td>
<td></td>
</tr>
<tr>
<td>Diabetes</td>
<td></td>
<td>V</td>
</tr>
</tbody>
</table>

Insurance premium for the children may be increased because lung cancer is suspect to genetic transmission.
The Dark Side (contd)

- Discovery of sensitive models

90% of all PhD students don’t do research!
The Nuclear Power Equivalence

How do we get all the good without suffering from the bad?
Our Focus

Addressing privacy concerns in the context of *Boolean Association Rule Mining*
Association Rules

• Co-occurrence of events:
  ➢ On supermarket purchases, indicates which items are typically bought together

  80 percent of customers purchasing coffee also purchased milk.

  $\text{Coffee } \Rightarrow \text{ Milk } \ (0.8)$

  To ensure statistical significance, need to also compute the “support” – coffee and milk are purchased together by 60 percent of customers.

  $\text{Coffee } \Rightarrow \text{ Milk } \ (0.8,0.6)$
Frequent Itemsets

- $T = \text{set of transactions}$
- $I = \text{set of items}$
- $sup_{\text{min}} = \text{User-specified threshold}$

“$X \subseteq I$ is frequent if more than $sup_{\text{min}}$ transactions in $T$, support $X$”
Privacy and BAR Mining

- Preventing discovery of sensitive rules
  - Atallah et al [KDEX 1999]
  - Saygin, Verykios, Clifton [SIGMOD Record 2001]
  - Dasseni, Verykios [IHW 2001]
  - Saygin et al [RIDE 2002]

- Preventing disclosure of data
  - Our work
  - Concurrent work by Evfimievski et al [KDD 2002]
Requirements for Mining with Data Privacy

• High Privacy
  ➢ User-visibility of privacy
• Highly accurate models
• Efficiency
  ➢ Data aggregation-time efficiency
  ➢ Mining-time efficiency
Conflicting Goals

Data Privacy  Accurate Models

Vs.
The Game Plan

User Data → Distorted Data

A Distortion Procedure

A Reconstruction Procedure

Pretty Accurate Models

Our Algorithm
Outline

- Privacy by data distortion
- Mining the distorted database (MASK)
- Experimental Evaluation
- Run-time Optimizations
- Conclusions, Limitations and Future Work
Distortion Procedure

• View the database as a matrix of 0s and 1s
  ➢ 0s represent absence of the item in the transaction
  ➢ 1s represent presence of the item in the transaction

Global data swapping? (privacy not “user-visible”)

Data perturbation?

• Independently flip some entries in the matrix. Don’t flip with probability \( p \), flip with probability \( 1-p \) (\( p=0.1 \) – 90% flips)
## Torvald’s Dilemma

### Original Customer Tuple

<table>
<thead>
<tr>
<th>Diapers</th>
<th>Insulin</th>
<th>Diet Coke</th>
<th>MS Office</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

1 = bought  
0 = not bought

### Distorted Tuple

<table>
<thead>
<tr>
<th>Diapers</th>
<th>Insulin</th>
<th>Diet Coke</th>
<th>MS Office</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Privacy Breach Measure

- Reconstruction probability of a ‘1’ in the $i^{th}$ column

$$P_{r\{Y_i=1|X_i=1\}} \times P_{r\{X_i=1|Y_i=1\}} + \quad \text{RECONSTRUCTED}$$

$$+ P_{r\{Y_i=0|X_i=1\}} \times P_{r\{X_i=1|Y_i=0\}}$$

$$\text{ORIGINAL}$$

$$\begin{array}{ccc}
X_i & 1 & 0 & 0 & 1 \\
Y_i & 1 & . & . & . \\
\end{array}$$

$$\begin{array}{ccc}
X_i & 0 & . & . & . \\
Y_i & ? & . & . & . \\
\end{array}$$

$$\text{DISTORTED}$$
Reconstruction Probability of a ‘1’

\[ R(p, s_i) = \frac{s_i p^2}{s_i p + (1-s_i)(1-p)} + \frac{s_i (1-p)^2}{s_i (1-p) + (1-s_i) p} \]

- \( s_i = \) support for item \( i \)
- \( p = \) distortion parameter

Graph showing \( R(p, s_i) \) for given \( s_i \)
Privacy Measure

\[ P(p, s_i) = (1 - R(p, s_i)) \times 100 \]

The Playground!

\[ P(p, s_i) \text{ for } s_i = 0.01 \]
Data Distortion and Psychology

<table>
<thead>
<tr>
<th></th>
<th>diapers</th>
<th>Insulin</th>
<th>Diet Coke</th>
<th>MS Office</th>
<th>..........</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>..........</td>
</tr>
</tbody>
</table>

$p=0.1$

0 0 1 0 ......

90% distortion

$p=0.9$

1 1 1 1 1 ......

10% distortion

More visible distortion $\Rightarrow$ Happier Customer?
Outline

- Privacy by data distortion
- Mining the distorted database (MASK)
- Experimental Evaluation
- Run-time Optimizations
- Conclusions, Limitations and Future Work
MASK
(Mining Associations with Secrecy Konstraints)

1. $F=\ ?$
2. $Cands=$Set of all items
3. $Length=1$
4. While $Cands$ ? 
   1. Count $2^{Length}$ components for each c ? $Cands$
   2. Reconstruct the support for each c ? $Cands$
   3. Add all frequent itemsets to F
   4. $Cands=$Apriori-Gen($Cands$)
   5. $Length=Length+1$
5. Return F
Counters

- $2^n$ counters for an $n$-itemset
- $\{c_{00}, c_{01}, c_{10}, c_{11}\}$ for a 2-itemset
- $\{c_{000}, c_{001}, c_{010}, c_{011}, c_{100}, c_{101}, c_{110}, c_{111}\}$ for a 3-itemset
MASK
(Mining Associations with Secrecy Konstraints)

1. $F=?$
2. $Cands =$ Set of all items
3. $Length = 1$
4. **While** $Cands$ ?
   1. Count $2^{Length}$ components for each $c$ ? $Cands$
   2. Reconstruct the support for each $c$ ? $Cands$
   3. Add all frequent itemsets to $F$
   4. $Cands = Apriori-Gen(Cands)$
   5. $Length = Length + 1$
5. **Return** $F$
Support Reconstruction for 1-itemsets

\[
\begin{bmatrix}
P & 1-P \\
1-P & P
\end{bmatrix}
\begin{bmatrix}
c_1 \\
c_0
\end{bmatrix}
= \begin{bmatrix}
c_D^1 \\
c_D^0
\end{bmatrix}
\]

\[
c_0, c_1 = 0,1 \text{ counts in the original column}
\]

\[
c_D^0, c_D^1 = 0,1 \text{ counts in the distorted column}
\]

\[
p = \text{distortion parameter}
\]

\[
C = M^{-1}C^D
\]
Support Reconstruction for an $n$-itemset

\[ C = M^{-1}C^D \]

- $C$ = Original $2^n$ Counts
- $C^D$ = Distorted $2^n$ Counts
  (e.g. counts for 00, 01, 10, 11 for a 2-itemset)

\[ M = \{m_{i,j}\} \]

- $m_{i,j} =$ probability that a tuple of the form $j$ distorts to a tuple of the form $i$
  - e.g. $m_{1,2}$ for a 3-itemset is the probability that a “010” tuple distorts to a “001” = $p \times (1-p) \times (1-p)$
The Big Picture

- User-visible Privacy
- Value of $p$ is pre-decided
- Data-miner gets both the distorted data and $p$
- Reconstruction of supports
Outline

- Privacy by data distortion
- Mining the distorted database (MASK)
- **Experimental Evaluation**
- Run-time Optimizations
- Conclusions, Limitations and Future Work
Error Metrics

- **Support Error**
  \[
  \rho = \frac{1}{|F|} \sum_f \frac{|rec\_sup_f - act\_sup_f|}{act\_sup_f} \times 100
  \]

- **Identity Error**
  \[
  \sigma^+ = \frac{|R - F|}{|F|} \times 100
  \]
  \[
  \sigma^- = \frac{|F - R|}{|F|} \times 100
  \]
  (false positives)
  (false negatives)

\(R\)=reconstructed set of frequent itemsets
\(F\)=actual set of frequent itemsets
The Setup

- **Scaled Real Dataset (BMS-WebView)**
  - 500 items
  - 0.6 million tuples

- **Synthetic Dataset (IBM Almaden)**
  - 1000 items
  - 1 million tuples

- Experiments across $p$ & $sup_{min}$ values

- Low $sup_{min}$ values are tough nuts
Results with $p=0.9$, $\text{sup}_{\text{min}}=0.25\%$

| Level | $|F|$ | $?$ | $s^-$ | $s^+$ |
|-------|------|-----|------|------|
| 1     | 249  | 5.9 | 4.0  | 2.8  |
| 2     | 239  | 3.9 | 6.7  | 7.1  |
| 3     | 73   | 2.6 | 11.0 | 9.6  |
| 4     | 4    | 1.4 | 0    | 25.0 |
### Results with $p=0.7$, $sup_{min}=0.25\%$

| Level | $|F|$ | $?$ | $s^-$ | $s^+$ |
|-------|------|-----|-------|-------|
| 1     | 249  | 19.0| 7.2   | 15.7  |
| 2     | 239  | 33.6| 20.1  | 1907.5|
| 3     | 73   | 32.9| 30.1  | 2308.2|
| 4     | 4    | 7.6 | 50.0  | 400.0 |
Effect of Relaxation
\( p=0.9, \text{ sup}_{\text{min}}=0.25\% \)

- 10% relaxation in \( \text{sup}_{\text{min}} \)

<table>
<thead>
<tr>
<th>Level</th>
<th></th>
<th>F</th>
<th></th>
<th>?</th>
<th>( s^- )</th>
<th>( s^+ )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>249</td>
<td>6.1</td>
<td>1.2</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>239</td>
<td>4.0</td>
<td>1.3</td>
<td>23.4</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>73</td>
<td>2.9</td>
<td>0</td>
<td>45.2</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>4</td>
<td>1.4</td>
<td>0</td>
<td>75.0</td>
<td></td>
</tr>
</tbody>
</table>
Summary of Experiments

- “Window of opportunity”: around $p=0.9$ (symmetrically $0.1$)
- Unusable Models as $p > 0.5$
- Significant loss of privacy as $p > 1, 0$
- Most identity errors occur near the $\text{sup}_{\text{min}}$ boundary
- Low errors at higher levels
Outline

- Distortion and Reconstruction
- Privacy Metric
- MASK Algorithm
- Experimental Evaluation
- Run-time Optimizations
- Conclusions, Limitations and Future Work
Linear Number of Counters

- Each row of $M_{2^n \times 2^n}$ in $C = M^{-1}C^D$ has only $n+1$ distinct entries.

- Example ($n = 2$):

  \[
  \text{count}(11) = a_0\text{count}^D(00) + a_1\text{count}^D(01) + a_2\text{count}^D(10) + a_3\text{count}^D(11)
  \]

  \[a_1 = a_2\]

- Only $n+1$ counters for an $n$-itemset.
Cutting Down on Counting

Example (pass 2):

- \( \text{count}^D(00) + \text{count}^D(01) + \text{count}^D(10) + \text{count}^D(11) = \text{dbsize} \)
- Disregard ‘00’ counts – since 01, 10 and 11 are already being counted
- Speeds up pass 2 in experimental runs (p~0.9) by a factor of 4
Outline

- Distortion and Reconstruction
- Privacy Metric
- MASK Algorithm
- Experimental Evaluation
- Run-time Optimizations
- **Conclusions, Limitations and Future Work**
Conclusions

- Simple probabilistic distortion of data: “User-visible”
- Achieves **conflicting goals** of privacy and model accuracy
- Optimizations **significantly reduce** time and space complexity
Limitations

- Even with the optimizations, the time complexity is high compared to standard (non-privacy-preserving) mining
- Does not take into account the re-interrogation of data with mining results [KDD02]
Future Work

- Improvements in running time
- Refinement of privacy estimates
- Extensions to generalized and quantitative association rules
Take Away

Like Reagan to Gorbachev on monitoring nuclear reductions: "Trust but verify", our motto is

"Trust, but distort"