Data Mining NetFlow
So What’s Next?

Mark E Kane
FloCon 2005
20 September 05
Objectives

- Data Mining, very briefly
- Frequency Patterns
- Discoveries
- Realizations
- Changes Made
Data Mining – automated extraction of previously unknown data that is interesting and potentially useful.
<table>
<thead>
<tr>
<th>Reality</th>
<th>Result of Data Mining</th>
<th>Example Analyst Hours</th>
<th>Example Investigator Hours</th>
<th>Example SysAdmin Hours</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>YES</td>
<td>YES</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>Crime Prevented / Prosecuted</td>
</tr>
<tr>
<td>NO</td>
<td>NO</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>YES</td>
<td>NO</td>
<td>∞</td>
<td>∞</td>
<td>∞</td>
<td>Time Lost to Investigate and Clean Up After Crime</td>
</tr>
<tr>
<td>NO</td>
<td>YES</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>Red Haring</td>
</tr>
</tbody>
</table>
Complexity of Mining NetFlow

- Shear Volume
- Complex Protocol Analysis
- Ambiguous Interpretations
- Very Smart Adversaries
Common Investigator Issues

- Undermanned and overworked
- Varied knowledge base
- Does not own networks
- No direct reporting structure
Data Mining Techniques

**Primary Techniques**
- Rule and Tree Induction
- Characterization
- **Classification**
- Regression
- Association
- **Clustering**

**Other Techniques**
- Dependency Modeling
- **Change Detection**
- Trend Analysis
- Deviation Detection
- **Link Analysis**
- Pattern Analysis
- Spatiotemporal Data Mining
- Mining Path Traversal Patterns
- **Mining Sequential/Frequent Patterns**

**Uncertain Reasoning Techniques**
- Fuzzy Logic
- Neural Networks
- Bayesian Networks
- Genetic Algorithms
- Rough Set Theory
Frequency Patterns

Mining Frequent Patterns in Data Streams in Multiple Time Granularities (Giennella, Han, Pei, Yan, and Yu)

- Support Decision Making
- Past Less Significant than Present
- Record Reduction
- Time Tilted Windows
### Interpreting Time-Tilted Windows

<table>
<thead>
<tr>
<th>DAY</th>
<th>Window Size</th>
<th>Transition</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Window</td>
<td>Transition</td>
<td>Size</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>N</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Y</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Y</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Y</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Monday</td>
<td>9</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tuesday</td>
<td>15</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wednesday</td>
<td>6</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thursday</td>
<td>6</td>
<td>6</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Friday</td>
<td>12</td>
<td>6</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Saturday</td>
<td>16</td>
<td>12</td>
<td>6</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sunday</td>
<td>6</td>
<td>14</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Monday</td>
<td>12</td>
<td>6</td>
<td>14</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tuesday</td>
<td>15</td>
<td>9</td>
<td>14</td>
<td>9</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Day 1: 9 events
Day 2: 15 events (two buckets)
Day 3: 6 events (two buckets)
Day 4: 6 events (two buckets)
Day 5: 16 events (three buckets)
Day 6: 12 events (four buckets)
## Presenting Frequency Patterns

<table>
<thead>
<tr>
<th>Window Number</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transaction 5</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Packet Support</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>16</td>
<td>32</td>
<td>32</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>8</td>
<td>16</td>
</tr>
<tr>
<td>Packet Size (Bytes)</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>16</td>
<td>32</td>
<td>32</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>8</td>
<td>16</td>
</tr>
</tbody>
</table>

### Table Data:

- **Presenting Frequency Patterns**
- **Window Number**:
  - 522.76: 2.13
  - 86.9: 1.26
  - 162.12: 0.76
- **Transaction Support**:
  - 2.53: 0.03
  - 2.22: 0.03
  - 1.86: 0.03
- **Packet Support**:
  - 0.31: 0.03
  - 0.28: 0.03
  - 0.26: 0.03

### Additional Notes:

- **Most Adverent IPs**
- **Frequency Patterns**
- **Notified Last Week**: 4
Failed email servers
Previously, unknown trusted relationships
Encryption without authentication
Possible, but unproven intrusions
Data Mining Results

Frustrated Investigators
Frustrated Analysts
One Very Frustrated Developer
Changes to Employ Data Mining

Establish common basis of understanding
Establish criteria for reporting
  - Geo-Resolution
  - Timeliness
  - Volume
Establish reporting procedures
Mark Kane

mkane @ ddktechgroup.com