Measurement Activity in CSTNET

Haina Tang
Outline

• introduction
• Network measurement and Monitoring
  – User behavior analysis
  – Anomaly Detection
• Challenge
introduction

• Netflow measurement
• Syslog
• Qos measurement
• High speed network measurement
• Integrated platform for network monitoring
• ....
Introduction (cont.)

- What I will focus on……
  - User behavior
  - Anomaly Detection
Motivation

• Know your users’ behavior
• Which scientific area is the most active one on your network?
• Security Considerations
System structures

- Netflow export layer.
- Data collecting and analysis layer.
  - this layer would filter, aggregate and store the result.
- Application analysis layer.
  - all sorts of traffic statistical comparison and analysis would be provided.
Pipelining for Dealing with the Data

- a design close to Pipelining is introduced to analyze the great capacity of data packets.
- let the runtime of the process \( i \) be \( \triangle t_i \), then the throughput is \( n/\sum \triangle t_i \) by the serial processing; or is the following formula by the parallel processing.

\[
TP = \frac{n}{T_n} \quad TP_{\text{max}} = \frac{1}{\lim_{n \to \infty} \max(\Delta t_1, \Delta t_2, \ldots, \Delta t_k)}
\]

- At the same time, optimizing these processes which restrict the system efficiency will improve the whole data dealing rate.
Analysis of User Behavior

• By analyzing the traffic of GLORIAD, detailed information of this international channel is shown by web service.

• All sorts of live (and historical) traffic analysis between world regions, countries, country regions, institutions, organizations, network protocols by year, month, day, hour, minute. . .
Our goal

• Traffic consumption
  – Which institute
  – Which country
  – AS*AS
  – BGP prefix
Which institute, country or region does one ip address belong to?

• The GeoIP Country database.
  – The database provides the mapping between one IP and its geographic address. The mapping is faster than WHOIS system and the accuracy is over 97%.

• IP address database
  – In order to acquire the institute to which one IP belongs, the Patricia Trie algorithm is used. In our testing, finding the IP institute information for 300000 records only costs 1 or 2 seconds.
Effectual Aggregation for Great Capacity of Data Packets

- Different combination for Space and time granularity, storage time of Netflow data.

<table>
<thead>
<tr>
<th>Space granularity</th>
<th>Time granularity</th>
<th>storage time</th>
</tr>
</thead>
<tbody>
<tr>
<td>IP</td>
<td>5 minutes</td>
<td>24 hours</td>
</tr>
<tr>
<td>institute</td>
<td>5 minutes</td>
<td>24 hours</td>
</tr>
<tr>
<td></td>
<td>24 hours</td>
<td>N years</td>
</tr>
<tr>
<td>countries, country regions</td>
<td>5 minutes</td>
<td>24 hours</td>
</tr>
<tr>
<td></td>
<td>24 hours</td>
<td>N years</td>
</tr>
<tr>
<td>autonomy system</td>
<td>5 minutes</td>
<td>24 hours</td>
</tr>
<tr>
<td></td>
<td>24 hours</td>
<td>N years</td>
</tr>
</tbody>
</table>
Top china institute received traffic from GLORIAD

1. 上海天文台
2. 地震局
3. 中科院生物物理所
4. 中科院大气所
5. 中科院高能物理所
6. 中科院计算所
7. 中科院力学所
8. 中科院化学所
9. 中科院物理所
10. 中科院福建物质结构所
Top China institutes send traffic into GLORIAD

1. 中科院高能物理所
2. 中科院大气所
3. 中科院情报中心
4. 中科院生物物理所
5. 中科院上海有机化学所
6. 中科院地理科学与资源研究所
7. 航天部
8. 地震局
9. 成都生物所
10. 中科院自动化所
Traffic trend for institute

- Traffic trend would respectively be shown by pks, pps, and fps for institutes, nations and regions, autonomous system.
Anomaly Detection

• Signature detection
  ❌ Require known attack signatures

• Anomaly detection
  – Alarm if deviation from a normal behavior model
  – Define the normal behavior
    • Rule-based Approach
      ❌ Manual rule specification time consuming
    • Learning-based Approach
      ❌ High false positive rates
Our Goal

• Automatic detection of aggregated anomalous events in a domain
  – Require **no foreknowledge** of normal state changes
  – Require **no system-specific** knowledge
Attack Models

• Anomalous event
  – An unexpected state change close in time across multiple nodes in a network system (exclude one exactly attack which can’t cause this change, such as Trojan Horse, script virus, ip address spoofing.)

• Focus
  – Attacks/events that take place at multiple nodes at a time
  – Internet worms, virus
  – Administrative updates
Node State Representation

• Statistical data change
  – most anomaly result in statistical data change, especially the anomaly concerned by us.

• Node representation
  – Throughput
  – Connection Speed
  – Network Efficiency
  – TCP Retransmit Percentages
  – Average Round Trip Times
Windows for Correlation (Sliding Window)

- **Detection window**: period for anomaly detection
  - One day for current prototype
- **Comparison window**: period to look back for comparison
- **Correlation window**: detection window + comparison window
Anomaly Detection by Clustering

- Raise an alarm if detect a *new cluster*
  - *New cluster*: consist of multiple vectors from *only* the detection window

- Regard normal if detect a *old cluster*
  - *Old cluster*: consist of multiple vectors from *both* the detection window and the comparison window
Anomaly Example

- Current data
- History data

input axis

output axis

New Cluster
Normal Example
First step: traffic model
Second step: traffic analysis
Third step: find the source

- the destination address 72.20.18.100 almost occupied all the flows of 40 octets packets.
- But the source addresses distribution is uniform for the flows of 40 octets packets.
The destination addresses which result in the flood flows of 40 octets packets

<table>
<thead>
<tr>
<th>top10</th>
<th>dstadd</th>
<th>flows(%)</th>
<th>octets(%)</th>
<th>package(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>72.20.18.100</td>
<td>76.525</td>
<td>2.604</td>
<td>26.716</td>
</tr>
<tr>
<td>2</td>
<td>81.3.150.143</td>
<td>4.691</td>
<td>1.223</td>
<td>14.931</td>
</tr>
<tr>
<td>3</td>
<td>159.226.8.159</td>
<td>0.488</td>
<td>0.03</td>
<td>0.197</td>
</tr>
<tr>
<td>4</td>
<td>159.226.1.1</td>
<td>0.325</td>
<td>0.03</td>
<td>0.16</td>
</tr>
<tr>
<td>5</td>
<td>159.226.1.3</td>
<td>0.252</td>
<td>0.018</td>
<td>0.105</td>
</tr>
<tr>
<td>6</td>
<td>210.77.69.6</td>
<td>0.219</td>
<td>0.01</td>
<td>0.078</td>
</tr>
<tr>
<td>7</td>
<td>159.226.1.8</td>
<td>0.178</td>
<td>0.017</td>
<td>0.086</td>
</tr>
<tr>
<td>8</td>
<td>159.226.156.11</td>
<td>0.155</td>
<td>0.023</td>
<td>0.075</td>
</tr>
<tr>
<td>9</td>
<td>159.226.6.176</td>
<td>0.09</td>
<td>0.007</td>
<td>0.039</td>
</tr>
<tr>
<td>10</td>
<td>210.75.224.23</td>
<td>0.07</td>
<td>0.016</td>
<td>0.044</td>
</tr>
</tbody>
</table>
The source addresses which result in the flood flows of 40 octets packets

<table>
<thead>
<tr>
<th>top10</th>
<th>srcadd</th>
<th>flows(%)</th>
<th>octets(%)</th>
<th>packets(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>210.77.69.6</td>
<td>4.692</td>
<td>1.223</td>
<td>14.932</td>
</tr>
<tr>
<td>2</td>
<td>159.226.8.159</td>
<td>0.677</td>
<td>0.056</td>
<td>0.293</td>
</tr>
<tr>
<td>3</td>
<td>159.226.1.1</td>
<td>0.389</td>
<td>0.074</td>
<td>0.191</td>
</tr>
<tr>
<td>4</td>
<td>159.226.1.3</td>
<td>0.273</td>
<td>0.041</td>
<td>0.12</td>
</tr>
<tr>
<td>5</td>
<td>159.226.50.9</td>
<td>0.241</td>
<td>0.036</td>
<td>0.126</td>
</tr>
<tr>
<td>6</td>
<td>81.3.150.143</td>
<td>0.219</td>
<td>0.01</td>
<td>0.078</td>
</tr>
<tr>
<td>7</td>
<td>159.226.156.11</td>
<td>0.186</td>
<td>0.026</td>
<td>0.103</td>
</tr>
<tr>
<td>8</td>
<td>159.226.1.8</td>
<td>0.186</td>
<td>0.046</td>
<td>0.102</td>
</tr>
<tr>
<td>9</td>
<td>210.75.224.23</td>
<td>0.181</td>
<td>0.014</td>
<td>0.078</td>
</tr>
<tr>
<td>10</td>
<td>202.127.16.210</td>
<td>0.169</td>
<td>0.007</td>
<td>0.062</td>
</tr>
</tbody>
</table>
Integrated Platform for network measurement

• Though different measure methods are applied, a integrated platform for network measurement has been built.

• This platform would get the evaluation from all measurement interfaces and would automatically create corresponding trouble ticket.

• The NOC engineers need not to switch the different monitor systems and can get the global alert information at real time.
Challenge

- How to improve positive rates
- Netflow data does not contain application layer information
- Performance
Thank you