Identification of Generic Attributes of Skype Traffic with Machine Learning

Rozanna Jesudasan
Centre for Advanced Internet Architectures (CAIA)
Swinburne University of Technology

Project Outline

- Statistical Identification of different versions of Skype.
- Identification of Games, Gtalk and Skype
- Identify traffic characteristics so that it will not change from one version to the next
- Identification done by extracting features (attributes) and using machine learning
Why Classify Traffic?

- Lawful Interception of Traffic
- Identify types of encrypted traffic
  - Skype
    - Restrict access
    - Lawful interception
    - Cisco NBAR can only detect Skype version 1
  - Games
    - Quality of Service
    - The ANGEL project at CAIA- Identify interactive traffic to provide better performance

Why Classify Traffic?

- Classification used in market research
  - Research on who is using what...
  - Research to better understand what applications people are using on the internet
- Quality of Service
  - ISP’s want to provision for traffic to minimise delays or increase efficiency
Methods to Classify Traffic

- Well known port numbers
  - But now we use higher port numbers
- Deep Packet Inspection
  - Computationally intensive
  - Does not work well with encrypted traffic like Skype
- Traffic Characteristics
  - How traffic varies with different features
  - Finding features to separately classify traffic

Previous Work at CAIA

- Identification of Skype Traffic using machine learning
- Identification of Bit Torrent Traffic using machine learning
- Identifying characteristics of games traffic
- Faster identification of traffic using sub flow characteristics

2. Philip Branch, Jason But, Tung Le, Rapid Identification of BitTorrent Traffic, 33rd Annual IEEE Conference on Local Computer Networks (LCN 2010), Denver, Colorado, USA, 11-14 October 2010
Previous Work at CAIA

- Work done previously relies mainly on packet lengths
  - But this characteristic usually changes from one version to another version
  - For Games it changes with the number of players

- Classification with Machine Learning
  - Classification with full flow and partial flows
  - Partial flows are quicker—especially for real-time traffic

Skype

- We examined Skype versions 2, 3, and 4
- Skype is a peer-to-peer VOIP application
- Skype can negotiate dynamic ports and it uses encrypted traffic
- The features across all versions differ
  - This is due to Skype using different codec's with each version
- We wanted to find a generic Skype classifier
Games

- Games characteristics change with increasing number of players
- From Previous work done at CAIA
  - Mean of Games Traffic increases linearly with increasing players
  - Variance of Games increases linearly with increasing players

Features

Features are statistical characteristics that can be used to identify different types of traffic.

- Mean Packet Payload
- Variance of Payload
- Ratio of Payload for Forward and Reverse Flows
- Inter-arrival time
- Index of Dispersion
- Two Packet Difference
- Absolute Two Packet Difference
Machine Learning

- A way to classify a given set of data using computer algorithms
  - Use the features that can differentiate each type of traffic to train the classifier.
  - After training it is tested against another set of data.

- Training
  - Train Skype V3 and “Other”
  - Training all Skype versions, Gtalk, Games and “Other”

- Testing
  - Test with Skype 4 and “Other”
  - Include Games into Other

Two classification algorithms used

- J48 classifier- Uses a decision tree and is a supervised learning classifier
- Naïve Bayes – A probabilistic classifier which is used for supervised learning
- We discovered that the J48 classifier performed better with differentiating between traffic types.
Tools

- Tcpdump- Capture traffic flows
- WEKA – classify traffic types
- Scripts – written by Dr. Jason But
  - Calculation of Features for sub flows, from a given pcap file.
  - Plotting CDF of Features
  - Converting .CSV to a .arff file (.arff in the WEKA input file with selected features)

Identification of Features

- The statistical characteristics calculated for the bidirectional, forward and reverse direction for each sub flow.
- Sub flows sized from 100 to 800 packets
- The CDF was plotted for all of the features for each traffic type.
- Features that were chosen
  - Mean, Two Packet Difference, Absolute Two Packet Difference, Index of Dispersion and Inter-arrival time.
Train on Skype V3 Test on Skype V4

- Skype Classification
  - Identify Skype version 4 traffic when trained by Skype version 3
- WEKA is given the Skype v3 and Other traffic
- Tested on Skype v4 for each combination of selected features
- Attributes that were successful in classifying
  - Mean, Index of Dispersion, Inter-arrival time and Two Packet Difference

Mean

**CDF**

- Skypev4
- Skypev3
- Skypev2
- Gtalk
- Other
- Games2
- Games4
- Games9

Subflow Size 100
Index of Dispersion

\[ D = \frac{\sigma^2}{\mu} \]

Two Packet Difference
Absolute Two Packet Difference

Inter-Arrival Time
Results

- Results shown are for classifying Skype for a sub-flow size of 100
- Results were obtained using J48 classifier

<table>
<thead>
<tr>
<th>Precision</th>
<th>Recall</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.974</td>
<td>0.858</td>
<td>Skype</td>
</tr>
<tr>
<td>0.979</td>
<td>0.996</td>
<td>Other</td>
</tr>
</tbody>
</table>

=== Confusion Matrix ===

a   b  <-- classified as
484 80 | a = Skype
13 3665 | b = Other

Conclusion

- Given the time frame, we were only able to train Skype, Games & Gtalk for certain attributes
- Training suggests that Skype and Games can be classified successfully
- Harder to classify Gtalk and Games
- Future work possible to create a more generic classifier
Acknowledgments

- I would like to thank Prof. Grenville Armitage for the opportunity to work at CAIA
- I would also like to thank Dr. Philip Branch and Dr. Jason But for their guidance and support
- I would like to thank Dr. David Hayes for his support.