Automated Traffic Classification and Application Identification using Machine Learning

Sebastian Zander, Thuy Nguyen, Grenville Armitage
{szander,tnguyen,garmitage}@swin.edu.au
Centre for Advanced Internet Architectures (CAIA)
Swinburne University of Technology

Supported by Cisco Systems, Inc. under the URP program

Outline

- Motivation
- Current Solutions & Shortfalls
- Machine Learning Approach
- Experimental Results
- Conclusions & Future Work

Motivation

- Different areas greatly benefit from classifying network traffic flows according to their creating applications
  - Application-based traffic trend analysis
  - Adaptive, network-based QoS mapping
  - Dynamic application-based access control
  - Lawful interception
  - Detection of malicious traffic

SWiN BUR NE CENTRE FOR ADVANCED INTERNET ARCHITECTURES

IEEE LCN 2005, Sydney, Australia, November 15th-17th http://caia.swin.edu.au szander@swin.edu.au Page 3

Current Solutions & Shortfalls 1/2

- Use port numbers for identification
  - Well-known and registered ports (IANA)
  - Known default ports (e.g. http://www.portsdb.org)
- Ambiguous default ports
- Applications use different or unknown ports
  - Multiple servers/clients on same IP address
  - Dynamically allocated ports (e.g. passive FTP)
  - Users deliberately using different ports (hide use of applications or bypass port-based filters)

SWiN BUR NE CENTRE FOR ADVANCED INTERNET ARCHITECTURES

IEEE LCN 2005, Sydney, Australia, November 15th-17th http://caia.swin.edu.au szander@swin.edu.au Page 4

Current Solutions & Shortfalls 2/2

- Stateful reconstruction of session and application information
  - Inspecting packet payload and decoding protocol
  - Resource intensive, must know the protocol (or reverse engineer), fails with encryption, privacy?
- Signature-based approach
  - Pattern search in packet payload
  - More efficient than protocol decoding but decreased accuracy, finding signatures can be difficult, fails with encryption, privacy?

SWIN BUR NE
CENTRE FOR ADVANCED INTERNET ARCHITECTURES

IEEE LCN 2005, Sydney, Australia, November 15th-17th http://caia.swin.edu.au szander@swin.edu.au Page 5

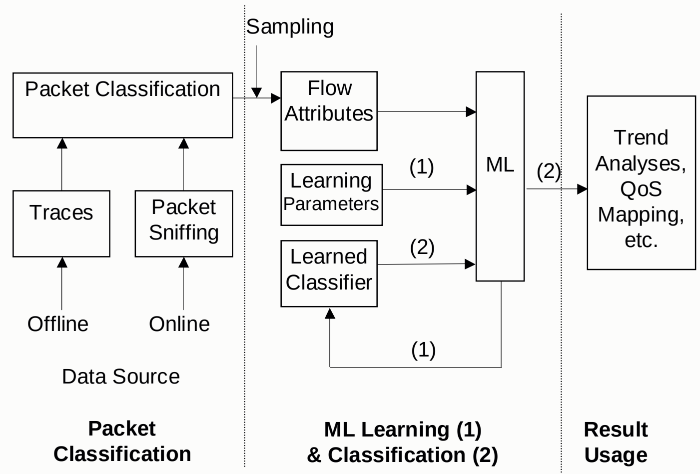
Machine Learning Approach 1/4

- Use protocol independent flow attributes (features)
  - Packet-level: e.g. packet length
  - Flow-level: e.g. inter-arrival times, duration, volume
  - Multi-flow-level: e.g. number of concurrent flows
- Use Machine Learning (ML) to classify flows using these features
  - Train algorithm on representative set of flows
  - Classify/predict classes for new unseen flows
- Idea is not completely new but lots of open questions
  - What algorithm? What (set of) features?
  - Accuracy? Performance?

SWIN BUR NE
CENTRE FOR ADVANCED INTERNET ARCHITECTURES

IEEE LCN 2005, Sydney, Australia, November 15th-17th http://caia.swin.edu.au szander@swin.edu.au Page 6

Machine Learning Approach 2/4



IEEE LCN 2005, Sydney, Australia, November 15th-17th http://caia.swin.edu.au szander@swin.edu.au Page 7

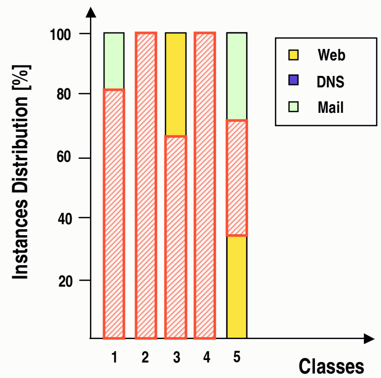
Machine Learning Approach 3/4

- Machine Learning Algorithm
  - Autoclass (http://ic.arc.nasa.gov/ic/projects/bayes-group/autoclass/)
  - Unsupervised learning (clustering)
- Feature selection
  - Sequential forward search (greedy algorithm)
    - Start with empty feature set
    - Each step add new feature that maximally increases goodness metric
  - Wrapper model (execute actual ML algorithm)
  - Goodness Metric: Intra-Class Homogeneity (H)
    - Percentage of instances of majority application in class

IEEE LCN 2005, Sydney, Australia, November 15th-17th http://caia.swin.edu.au szander@swin.edu.au Page 8

Machine Learning Approach 4/4

Example Homogeneity (H) computation

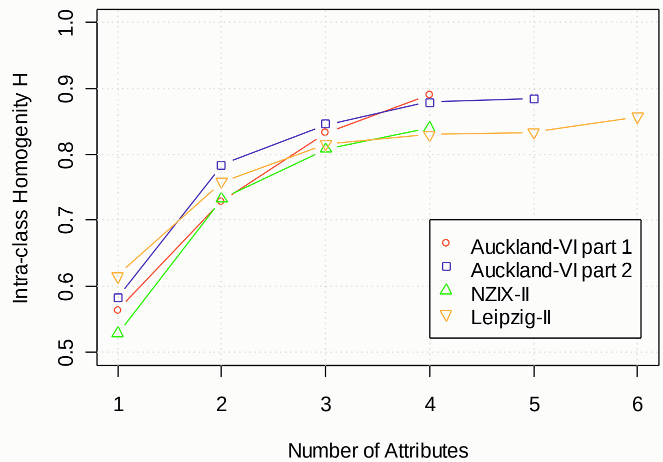


Class	App	H [%]
1	Web	82
2	DNS	100
3	Mail	67
4	Web	100
5	DNS	38
Total Average H		**77.4**

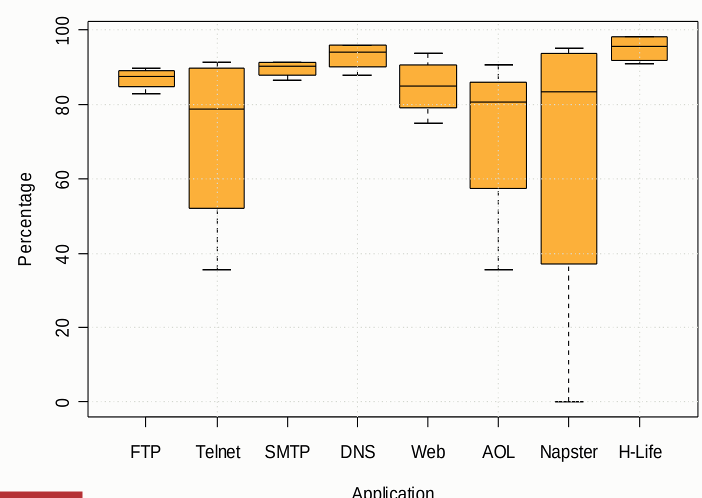
Dataset

- Packet traces from NLANR (http://www.nlanr.net)
  - Auckland VI (2 days), Leipzig II, NZIX II
  - 8 different applications: FTP Data, Telnet, Mail (SMTP), DNS, Web, AOL Messenger, Napster, Half-life
  - 1000 randomly sampled flows for each application
  - No payload in public traces
    - Select flows based on application default ports
    - Assume most flows are of expected application
    - Some ‘wrong’ flows decrease homogeneity
- Flow Attributes (Features)
  - Packet length (mean/variance), inter-arrival times (mean/variance), volume (bytes), duration
  - Bidirectional (except duration)

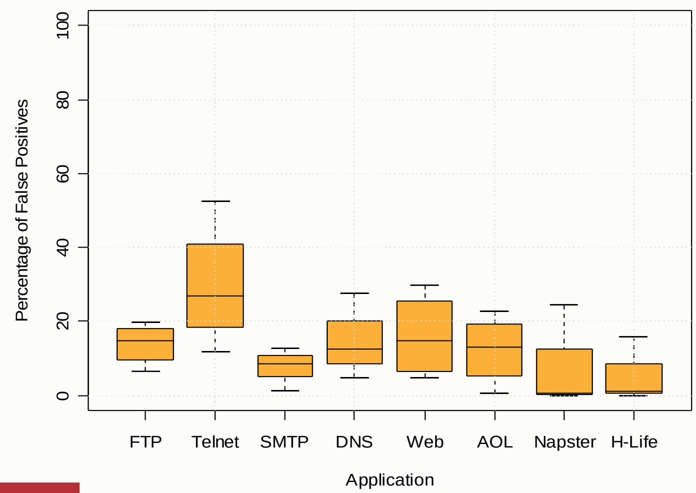
Results – Average Homogeneity



Results – Accuracy



Results – False Positives

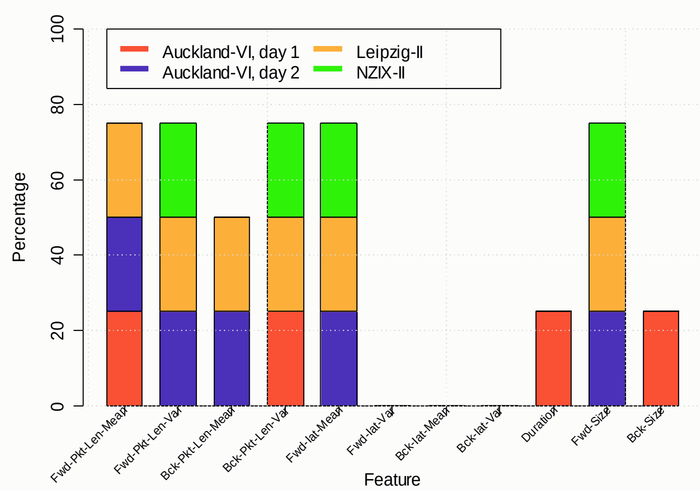


SWINBURNE
CENTRE FOR ADVANCED INTERNET ARCHITECTURES

IEEE LCN 2005, Sydney, Australia, November 15th-17th

http://caia.swin.edu.au szander@swin.edu.au Page 13

Results – Best Features



SWINBURNE
CENTRE FOR ADVANCED INTERNET ARCHITECTURES

IEEE LCN 2005, Sydney, Australia, November 15th-17th

http://caia.swin.edu.au szander@swin.edu.au Page 14

Conclusions

- Some separation of applications can be achieved
  - Average accuracy 86.5%
- Features
  - Packet length, volume favoured over inter-arrival times, duration (biased by our set of applications!)
- Performance (2.4GHz Celeron)
  - Learning very slow (~8.5 hours with full feature set)
  - Classification fast (~6,300 flows/second)
- Disadvantages of current ML technique
  - Classes need to be mapped to applications
  - Many parameters to be tuned

SWIN BUR NE — CENTRE FOR ADVANCED INTERNET ARCHITECTURES

IEEE LCN 2005, Sydney, Australia, November 15th-17th   http://caia.swin.edu.au szander@swin.edu.au   Page 15

Future Work

- Compared different ML algorithms (especially supervised techniques)
- Compare different feature selection methods
- Investigate new features
- For verification use traces where real application 'is known' (payload analysis)
- Investigate how quickly flows can be classified
- Investigate influence of flow sampling
- Investigate different application (e.g. peer-to-peer)
- Develop prototype software

SWIN BUR NE — CENTRE FOR ADVANCED INTERNET ARCHITECTURES

IEEE LCN 2005, Sydney, Australia, November 15th-17th   http://caia.swin.edu.au szander@swin.edu.au   Page 16

The End

Questions, Comments?