SWINBURNE

CENTRE FOR ADVANCED INTERNET ARCHITECTURES

Self-learning IP Traffic Classification based on Statistical Flow Characteristics

- Work in Progress

Sebastian Zander, Thuy Nguyen
{szander,tnguyen}@swin.edu.au

14/10/2004

Motivation

- Flow: set of IP packets passing an observation point in the network where all packets belonging to a particular flow have a set of common properties (e.g. packet header fields etc.). (RFC3917)
- Key areas in IP network engineering, management and surveillance that greatly benefit from classifying flows according to their responsible applications
  - Capacity demanding trend analysis
  - Application-based traffic engineering, monitoring
  - Adaptive, network-based QoS mapping
  - Dynamic application-based access control
  - Lawful interception
  - Intrusion detection

SWINBURNE CENTRE FOR ADVANCED INTERNET ARCHITECTURES

2

Current Solutions …

- Identify Applications by the destination port
  - Well-known and registered ports (assigned by IANA, e.g. /etc/services)
  - Known default ports (e.g. ports database: http://www.portsdb.org)
- Stateful reconstruction of session and application information
  - Inspecting packet contents and decoding the protocol
- Signature-based approach
  - Pattern search in the packet content

SWiN BUR NE CENTRE FOR ADVANCED INTERNET ARCHITECTURES

3

… and their Shortfalls

- Applications may use different or unknown destination ports
  - non-privileged users must run servers on ports higher >1024
  - users deliberately using different ports (hide the use of applications or bypass port-based filters)
  - multiple servers running on the same machine
  - dynamically allocated ports (passive FTP, streaming etc.)
- Stateful protocol decoding
  - Resource intensive and must know the protocol specification (or use reverse engineering)
  - Fails in case encryption is used
  - Privacy?
- Protocol fingerprinting
  - More efficient than protocol decoding
  - Decreased accuracy ('small' classification error)
  - Still requires knowledge about the protocol and no encryption
  - Privacy?

SWiN BUR NE CENTRE FOR ADVANCED INTERNET ARCHITECTURES

4

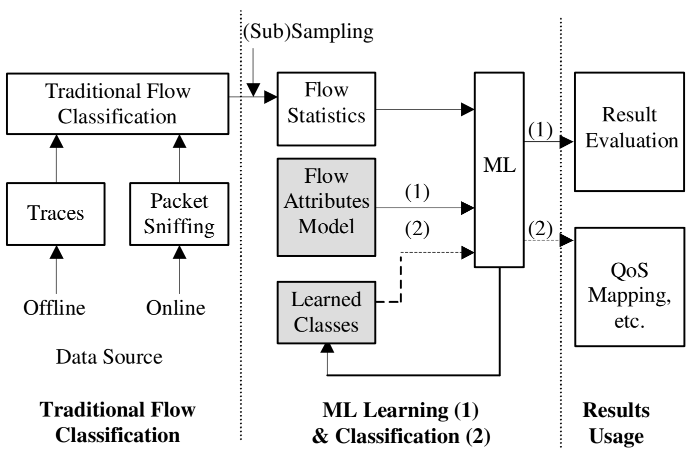
The "New" Approach

- Use flow attributes that can be derived without much effort and are protocol independent
  - Packet level: e.g. packet length
  - Flow level: e.g. inter-arrival times, duration
  - Inter-flow level: combination of different flows
- Use machine learning (ML) techniques to classify the flows according to their creating applications
  - Idea has been introduced some time ago in the security area
  - But still lots of open questions
- What (set of) flow attributes?
- What learning technique?
  - Supervised vs. unsupervised learning
- Accuracy? Performance? Usability?

SWINBURNE CENTRE FOR ADVANCED INTERNET ARCHITECTURES

5

Our Approach



SWINBURNE CENTRE FOR ADVANCED INTERNET ARCHITECTURES

6

Our Approach con't

- Traditional flow classification
  - Classification based on 5-tuple (IP addresses, ports and protocol) and computation of the flow statistics
  - Input: trace files (tcpdump format) or online capturing
  - Based on *NetMate* meter
- (Sub)sampling
  - Define maximum sample size n_{max}
  - If number of flows for an application is larger than n_{max} use random sampling for that application
  - Remove all flows that have less than 3 packets in one direction

SWiN BUR NE CENTRE FOR ADVANCED INTERNET ARCHITECTURES

7

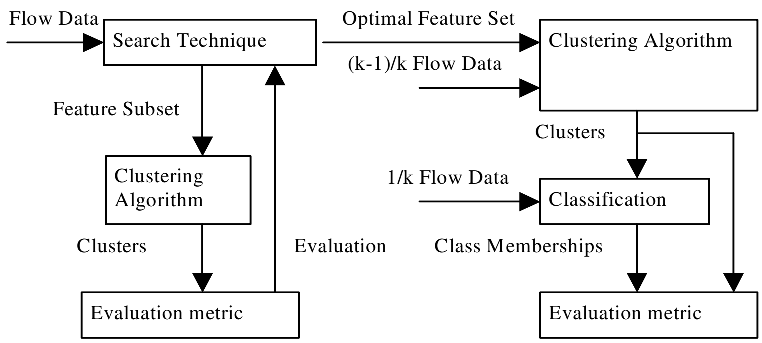
Our Approach con't

- Machine Learning (ML)
  - Unsupervised stochastic modeling: Expectation Maximization (EM) algorithm
  - 2 phases: learning/clustering and classification/prediction
  - Learning input: flow data, attribute definitions and models
  - Classification input: flow data, learned classes
  - Based on *autoclass*
- Evaluation/Result Usage
  - Evaluate Learning/classification accuracy
  - Apply classification results to QoS mapping etc.

SWiN BUR NE CENTRE FOR ADVANCED INTERNET ARCHITECTURES

8

Evaluation



Evaluation con't

- In case we don not the true application assume the real application is given by the destination port
- Find optimal attribute set
  - Linear incremental trial method
  - Metrics: Intra-class homogeneity (H) and inter-class spread (S)
  - Measure influence of attributes
- Evaluate the stability of the learned classes (attribute distribution, class size, number of classes
  - Cross validation with different traces
- Evaluate the classification accuracy
  - Cross validation with different traces
  - Metrics: Precision and recall
- Evaluate the flow sampling approach
- Evaluate the learning and classification performance

Preliminary Results

- 24 hour trace from the Auckland VI dataset (~90 million packet headers) contains 3-4 million flows
- Investigate 8 different applications (ports)
  - FTP Data
  - Telnet
  - SMTP
  - DNS
  - Web (port 80)
  - AOL Messenger
  - Napster
  - Half-life
- Use sample of 150 flows per application (=1200 flows)

SWiN BUR NE CENTRE FOR ADVANCED INTERNET ARCHITECTURES

11

Preliminary Results

- Initial trials with different flow attribute combinations
- All attributions have two directions

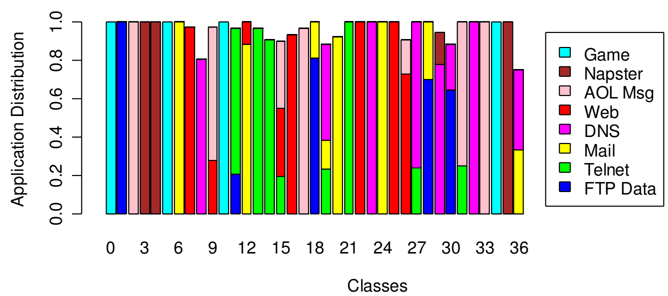
- Trial 1: mean packet length (PL)
- Trial 2: mean inter-arrival time (IAT)
- Trial 3: mean PL and mean IAT
- Trial 4: mean and std deviation of PL and IAT
- Trial 5: as Trial 4 plus flow duration and size (bytes)

SWiN BUR NE CENTRE FOR ADVANCED INTERNET ARCHITECTURES

12

Preliminary Results

	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5
H	0.78	0.48	0.77	**0.85**	0.85
S	0.49	0.76	0.35	**0.26**	0.27
C	18	23	41	**50**	57



SWINBURNE CENTRE FOR ADVANCED INTERNET ARCHITECTURES

13

Conclusions & Future Work

- Basic approach implemented and framework for evaluation defined
- Preliminary results show that some separation of the applications can be achieved

- Comprehensive evaluation based on longer traces
- Refining the approach
- Would be good to have traces where the real application is not only known by the port number
- Performance: need the Swinburne supercomputer ☺

SWINBURNE CENTRE FOR ADVANCED INTERNET ARCHITECTURES

14

The End

Thanks for your attention!
Questions, Comments?

SWINBURNE CENTRE FOR ADVANCED INTERNET ARCHITECTURES

15