An Automatic Traffic Classification System for Operational Networks

2nd TMA PhD School, Napoli 2011

Valentín Carela-Español
Pere Barlet-Ros
Josep Solé-Pareta

{vcarela, pbarlet, pareta}@ac.upc.edu
Outline

- Introduction
- Literature
- Proposed solution
  - The Application Identifier
  - The Automatic Retraining System
- Evaluation
Outline

- Introduction
- Literature
- Proposed solution
  - The Application Identifier
  - The Automatic Retraining System
- Evaluation
Introduction

● What is application identification?

● What is application identification used for?
  • Network planning and dimensioning
  • Performance evaluation
  • Charging and billing
  • QoS policies
  • Research purposes
Introduction

- Tons of papers have presented novel classification solutions

- Network operators still use obsolete methods

- What is slowing down the deployment of these novel techniques? Which are the features that operators are interested in?
  - Easy to deploy
  - Easy to maintain
  - High accuracy and completeness
Outline

- Introduction
- Literature
- Proposed solution
  - The Application Identifier
  - The Automatic Retraining System
- Evaluation
Well-known ports
  - Low accuracy and completeness

DPI-based (pattern matching)
  - Do not classify encrypted traffic
  - Computationally expensive

Machine Learning-based
  - Difficult training phase

Host-behaviour-based
  - Highly dependent on the monitoring point
  - Do not differentiate between applications with the same behaviour

Service and IP-based
  - Low completeness
Outline

- Introduction
- Literature
- Proposed solution
  - The Application Identifier
  - The Automatic Retraining System
- Evaluation
Proposed solution

How we deal with operators constraints?

- **Easy to deploy**
  - Using Sampled NetFlow as input for the classifier
    - NetFlow is already available on routers
    - Sampling decreases the load of the routers

- **Easy to maintain**
  - The automatic retraining system

- **High accuracy and completeness**
  - Combination of different techniques from the literature
The Application Identifier

- Using only Sampled NetFlow (v5) data as input
  - \(<\text{IPs, ports, \#pkts, \#bytes, protocol, ToS, TCP flags and duration}>\)

- Method based on multiple classification techniques:
  - Machine Learning-based [1]
    - Using the C5.0 technique, the enhanced successor of the well-known C4.5 decision tree, with the optimization proposed in [1] to deal with sampled traffic
  - Service-based [2]
    - Automatic detection of services: \(<\text{IP, port, protocol}>\) assigned to a specific application.
  - IP-based [3]
    - Using the IPs from well-known video-sharing applications (e.g. youtube, megavideo)

**The Automatic Retraining System**

- **Automatic Retraining System**
  - Continuously check the classifier accuracy
  - Retrain the system when the accuracy falls below a threshold

- Combines different DPI techniques to set the ground truth
  - PACE
  - OpenDPI
  - L7-Filter
Outline

- Introduction
- Literature
- Proposed solution
  - The Application Identifier
  - The Automatic Retraining System
- Evaluation
Evaluation

• Dataset
  • UPC-II trace [1] for the initial training of the classifier
    – A fifteen-minutes full payload trace collected in December 2008 at
      the Gigabit access link of the UPC.
  • CESCA trace for the validation
    – A fourteen-days packet trace collected in February 2011 at the
      10-Gigabit access link of the Anella Científica, link that connects
      the Catalan RREN with the Spanish NREN. Collected with a
      1/400 flow sampling rate.

• Retraining policy
  • Different accuracy thresholds (94%, 96%, 98%)
  • 500,000 flows as upper bound for the retraining
  • Flows from the last 7 days as base-truth.

Evaluation
Thank you for your attention

Questions?