Internet traffic classification based on flows' statistical properties with machine learning

Speaker: Lin Wang

Research Advisor: Biswanath Mukherjee

Vlăduţu, Alina, Dragoş Comăneci, and Ciprian Dobre. "Internet traffic classification based on flows' statistical properties with machine learning," International Journal of Network Management vol. 27, no. 3, 2017.

UCDAVIS

Motivation

- **Traffic demand increasing**

- **Scaling the network horizontally**
  - Improve network with more powerful machines
- **Scaling the network horizontally**
  - Bring more machines of the same power as the current ones

- **A new solution**
  - Apply machine learning techniques to classify network traffic in order to detect any traffic patterns and adjust the resources accordingly.

Introduction

- **State-of-art**
  - Classify traffic based on the packets' statistical properties

  - In this work
  - Extract statistical properties of the packets
  - Apply K-means to group them together in clusters based on similarities
  - Use this classification along with all the statistical properties to train a supervised learning engine using C4.5.
  - Classify new traffic using above well-trained C4.5.

Deep packet inspection (DPI)

- **Every packet should be checked against the available traffic signatures**
- **High accuracy**
- **A greedy resource consumer and not scalable**
- **Useless in case of encrypted traffic**
  - (i.e., protocols like SSH or HTTPS)
- **Network congestion results in latency**

Statistical properties of a flow

- duration of the flow
- total number of packets involved
- packets length taken individually or in total
- flow length in bytes
- inter-packet arrival time.

Unsupervised learning

- **No labeled input data and tries to find any hidden properties inside it.**

- **Why useful?**
  - underlying structure: to obtain an insight on how the data look like, to detect features or anomalies;
  - natural classification: to identify similarities between different organisms;
  - diversity of clusters: to identify groups based on different criteria;
  - compression: to organize data based on the cluster prototypes.

K-means Clustering

- **Flow**
  - A flow is a five-dimensional tuple: (*source IP, destination IP, source port, destination port, and transport protocol*).
- **Unidirectional flows**
  - Composed of packets that are going in one (from A to B) direction.
- **Bidirectional flows**
  - Composed of packets that are going in two (from A to B and back from B to A) directions.

Statistical properties for flows

- **Unidirectional Flow**
  - number of packets;
  - duration of the flow (the time between the last packet and the first packet sent);
  - total length of all/first 10 packets;
  - minimum/maximum/average/standard deviation packet length;
  - minimum/maximum/average/standard deviation inter-packet arrival time.

Statistical properties for flows

- **Bidirectional flows**
  - duration of the flow (with the same meaning as in the unidirectional flows);
  - first 10 senders (which peer sent each of the first 10 packets);
  - first 10 inter-packet arrival times;
  - first 10 packets length;
  - first 10 differences between inter-packet arrival time;
  - number of packets sent;
  - minimum/maximum/average/standard deviation packet length;
  - minimum/maximum/average/standard deviation inter-packet arrival time.

Types of traffic in percentage in our experimental capture

Traffic type	Percentage from the whole traffic
HTTP Video Enterprise	29.50
HTTP Enterprise	18.63
SSH Enterprise	17.96
Oracle Enterprise	17.15
Raw UDP Enterprise	5.96
Raw Enterprise	3.71
BitTorrent Enterprise	2.09
Flash Enterprise	1.93
HTTPS Simulated Enterprise	0.90
SMB Enterprise	0.75
SMTP Enterprise	0.56
PPLive Enterprise	0.51
FTP Enterprise	0.32
YouTube Enterprise	0.03

Experiment results

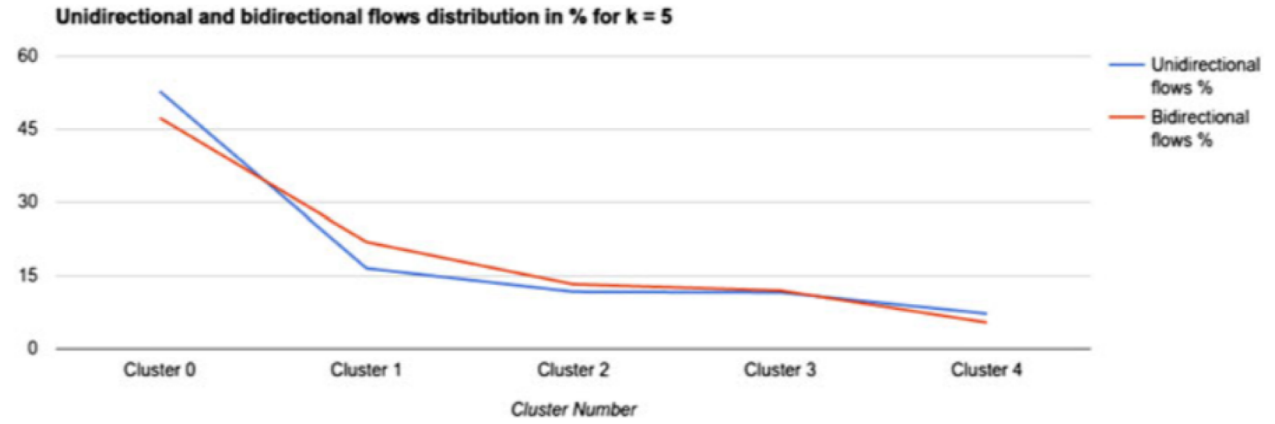


Figure 2. Unidirectional and bidirectional flow distribution in percentage for k = 5

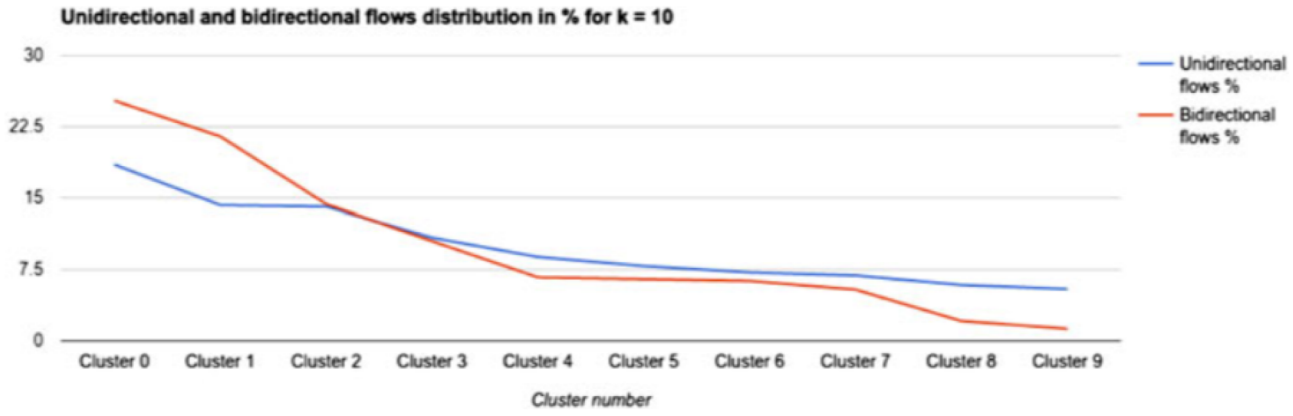


Figure 3. Unidirectional and bidirectional flow distribution in percentage for k = 10

Experiment results

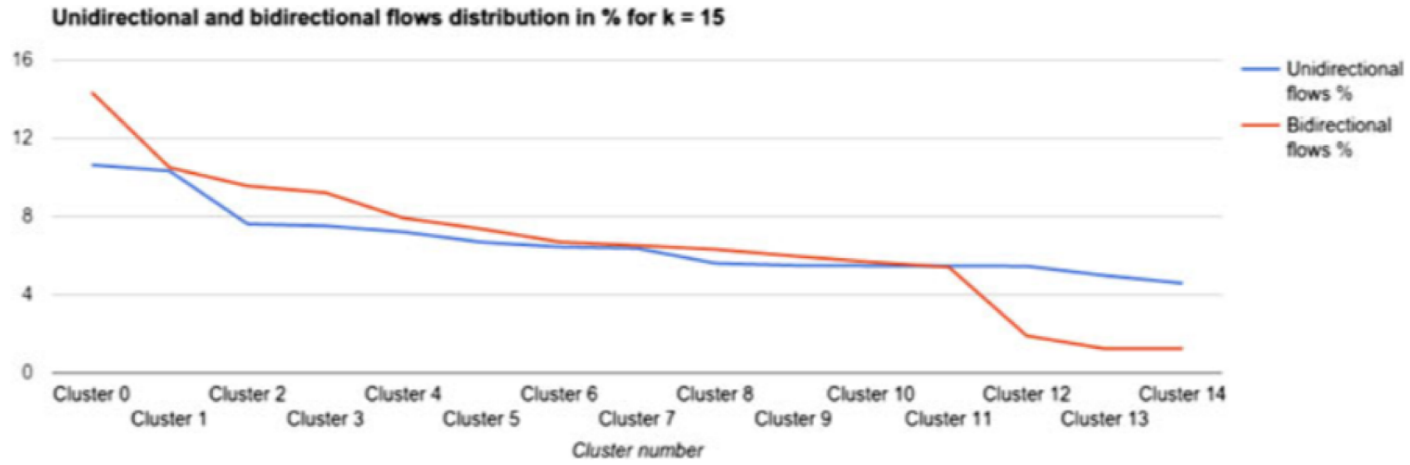


Figure 4. Unidirectional and bidirectional flow distribution in percentage for k = 15

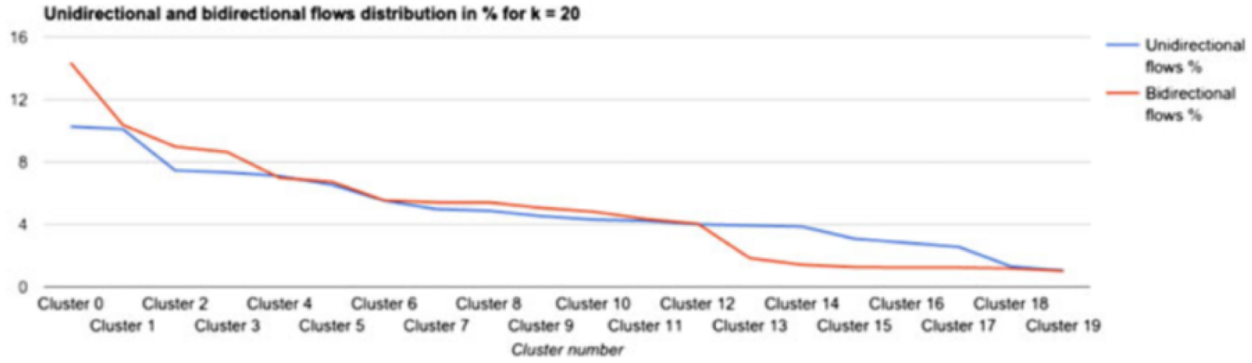
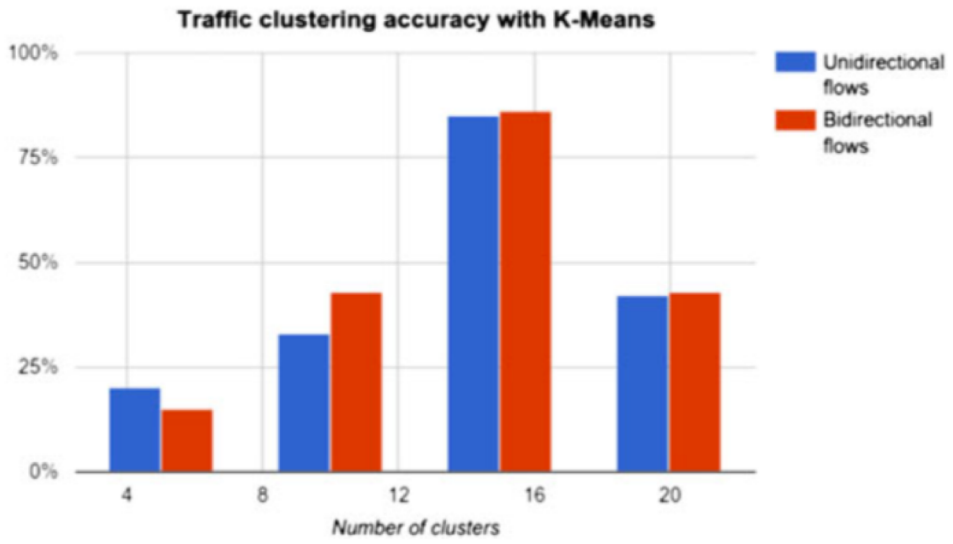


Figure 5. Unidirectional and bidirectional flows distribution in percentage for k = 20

Experiment results

Table 2. Clustering accuracy for unidirectional and bidirectional flows for multiple number of clusters

Number of clusters	Unidirectional flows	Bidirectional flows
5	0.20	0.15
10	0.33	0.43
15	0.85	0.86
20	0.42	0.43

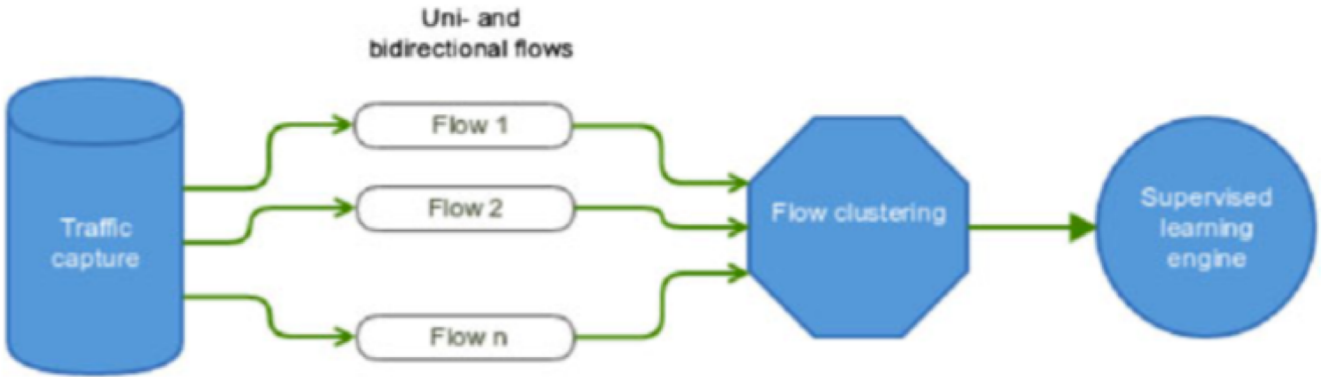


Supervised learning

- **Infer a function based on a pre-labeled training set.**
- **User well-trained model to classify new data.**

- **Why C4.5?**
  - Fast speed and low cost
  - accepts both discrete and continuous attributes
  - C4.5 accepts missing attribute values.

Project Design



- Use the output of the K-means (k=15) clusterization as input for a supervised learning
- Having 50%, 70%, and 90% of the input used for training and the other portion for testing,
- Over 90% of the flows were classified correctly.

UCDAVIS
UNIVERSITY OF CALIFORNIA

amlwang@ucdavis.edu