LANCASTER UNIVERSITY
Computing Department
Faculty of Applied Sciences

InfoLab21

Network Traffic Characterization using Energy TF Distributions

Angelos K. Marnerides

a.marnerides@comp.lancs.ac.uk

Collaborators:
David Hutchison - Lancaster University
Dimitrios P. Pezaros - University of Glasgow
Hyun-chul Kim -Seoul National University

University of Glasgow
Department of Computing Science

SEOUL NATIONAL UNIVERSITY

Outline

- Motivation
- Approach
- Data & Features
- Results
- Summary
- On-going & Future Work

Importance of Traffic Characterization & Classification

- Weakness of manual inspection by NOCs
- Pre-requisite for understanding the fluctuant network behavior
- Foundational element for Traffic Engineering (TE) tasks:

  - cost optimization ,efficient routing, congestion management, availability, resilience, anomaly detection, traffic classification etc..
- Application-based traffic Classification : a necessity

  - net neutrality debate, ISPs vs. Content providers
  - emergence of new applications, attacks etc..
  - file sharing vs. intellectual property representatives

Motivation

- Traffic modeling assumptions not thoroughly investigated
  - Stationarity?

- Rapid growth of new Internet technologies and applications.

- Essence for new and adaptive traffic classification features.

Approach

- Volume-based analysis on real pre-captured network traces for characterizing the traffic's dynamics.

- Validation of stationarity under TF representations
  - Instantaneous frequency and group delay for stationarity.

- Volume decomposition for revealing protocol-specific dynamics and classify the volume-wise utilization (#bytes and #pkts) of the transport layer.

- Provision of application-layer characteristics based on the level of signal complexity using the Cohen-based Energy TF Distributions.

Data & Features

- 2 30min full pcap traces from a Gb Ethernet Link at Keio University, Japan (Keio-I, Keio-II)

  - extracted # of bytes & pkts for each unidirectional flow for TCP,UDP, ICMP

- Hour-long full pcap trace from a US-JP link (WIDE) 100 Mbps FastEthernet link (SamplePoint B – MAWI Working group)

  - divided in 4, 13.75-min bins (WIDE-I,WIDE-II,WIDE-III,WIDE-IV)

  -employed the same feature extraction as in Keio-I/II

Data & Features (tables)

Table 1: Captured Operational Traces from WIDE & Keio

Set	Date	Day	Start	Duration	Link Type	Packets	Bytes	Avg. Util.	Flows/min
WIDE	03-03-2006	Fri	22:45	55min	Backbone	32M	14G	35Mbps	63K
Keio-I	06-08-2006	Tue	19:43	30min	Edge	27M	16G	75Mbps	32K
Keio-II	10-08-2006	Thu	01:18	30min	Edge	25M	16G	75Mbps	19K

Table 2: Traces pre-processing

Set	Duration	TCP flows/min	UDP flows/min	ICMP flows/min
WIDE-I	13.75min	24K	30K	4K
WIDE-II	13.75min	28K	31K	4K
WIDE-III	13.75min	24K	29K	3K
WIDE-IV	13.75min	23K	30K	4K
Keio-I	30min	21K	10K	6K
Keio-II	30min	8K	9K	4K

*** Kim et al. L., *Internet traffic classification demystified: myths, caveats, and the best practices*, ACM CoNEXT 2008**

Stationarity Test

- A signal is stationary if the elements in its analytical form keep a constant instantaneous frequency and group delay respectively.

Process *g(t)* (counts of bytes/packets), and $G_a(t)$ its analytical form after applying a Hilbert transformation and $F_a(\nu)$ the Fourier transform of $G_a(t)$

- Instantaneous Frequency → $f(t) = \frac{1}{2\pi} \frac{d \arg G_a(t)}{dt}$

  - *f(t):* amplitude of frequency we observe in 1 count of a packet/byte arrival at time *t*

- Group Delay → $t_G(\nu) = -\frac{1}{2\pi} \frac{d \arg F_a(\nu)}{d\nu}$

  - $t_G(\nu)$: time distortion caused by the signal's instantaneous frequency

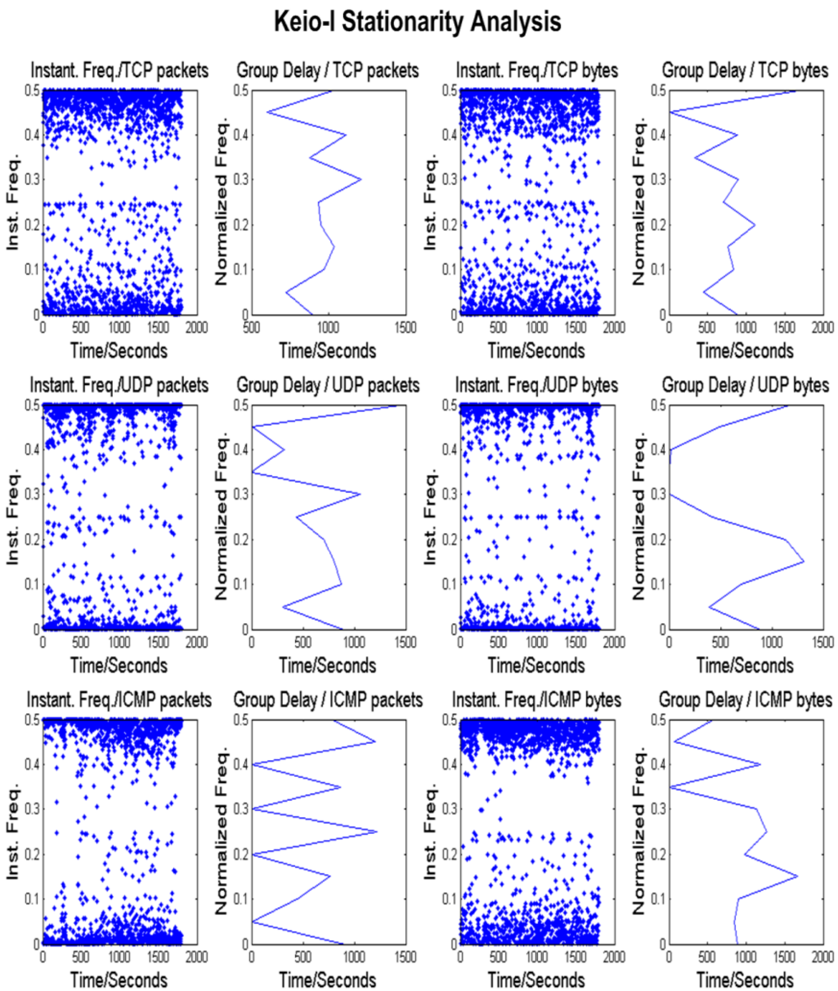
Stationarity analysis

- Validation of instantaneous frequency and group delay’s behaviour in all datasets.

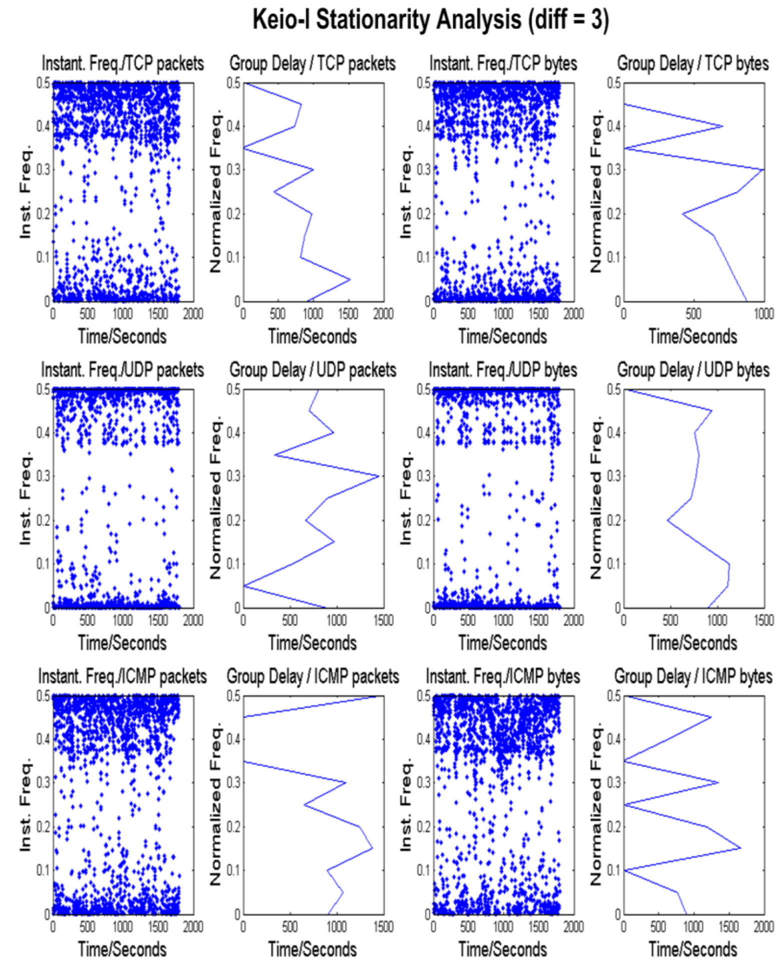
- Investigated stationarity on ithe original and differentiated traffic signal

- Conclusion : traffic in all traces is highly non-stationary and has the form of a multi-component signal (for all protocols).

Stationarity analysis (results)



Before differentiation



After 3rd order differentiation

Traffic Classification with Cohen-based Energy TF distributions

- Suitable for characterizing highly non-stationary signals as the volume dynamics of the transport layer.

  - Overcome limitations by other techniques (e.g. STFT, Wavelets) on the TF plane with respect to TF localization and resolution

- Particularly used *:

  -Wigner-Ville (WV) Distribution

  -Smoothed Pseudo Wigner-Ville (SPWV) Distribution

  - Choi-Williams (CW) Distribution

- Employment of Renyi Dimension for determining signal complexity (i.e. volume-wise intensity) on the TF plane – used as the classification discriminative feature

- Simple Decision tree-based classification using MATLAB's classification utility functions

Definitions provided in : Cohen, L., Time-Frequency Distributions: A Review, Proc IEEE Signal Processing, Vol. 77, 1989

Classification Performance Metrics

- Accuracy per-trace

$$Accuracy = \frac{\#correcty\_classified\_flows}{\#total\_flows\_per\_trace}$$

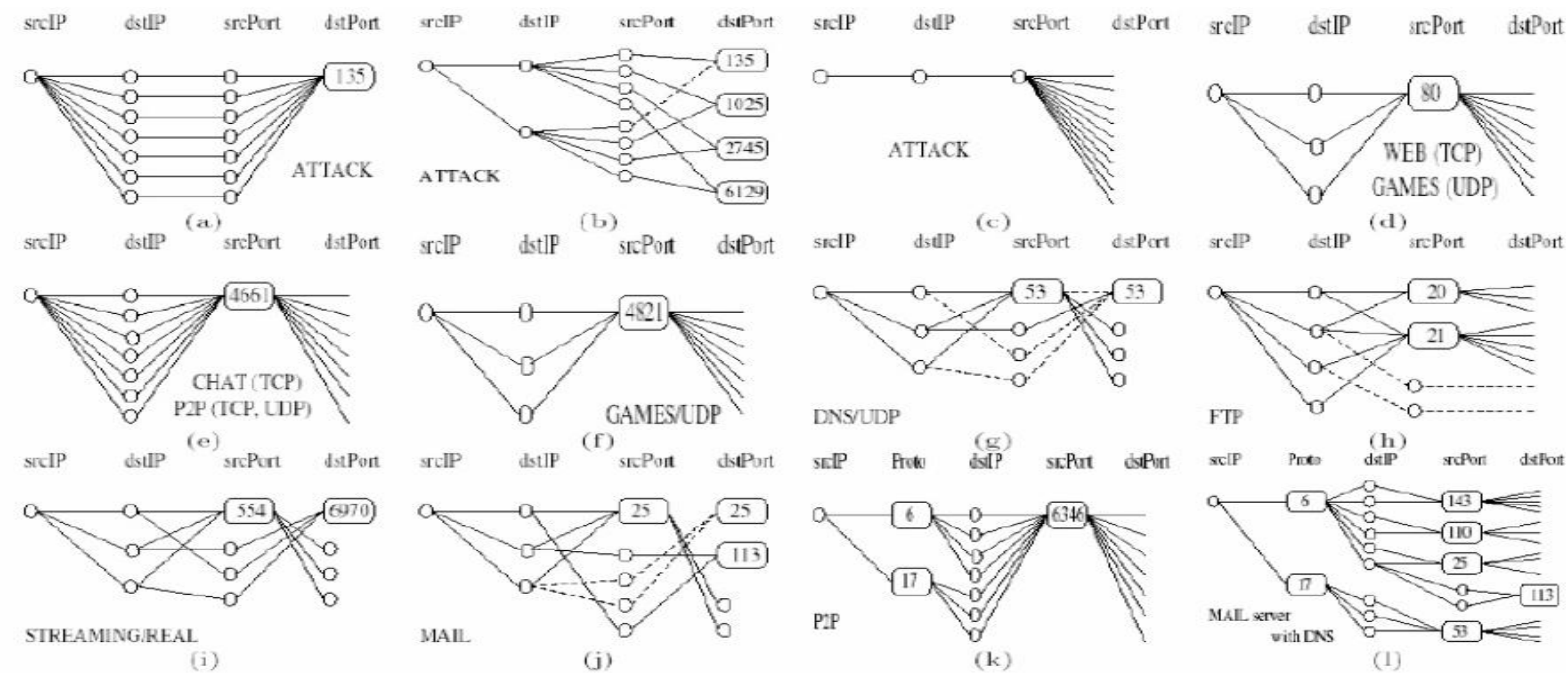
- Per-Application

- Recall : “How complete is an application fingerprint?”

$$\mathrm{Re}call = \frac{True\_positives}{True\_positives + False\_negatives}$$

Pre-processing for Traffic Classification

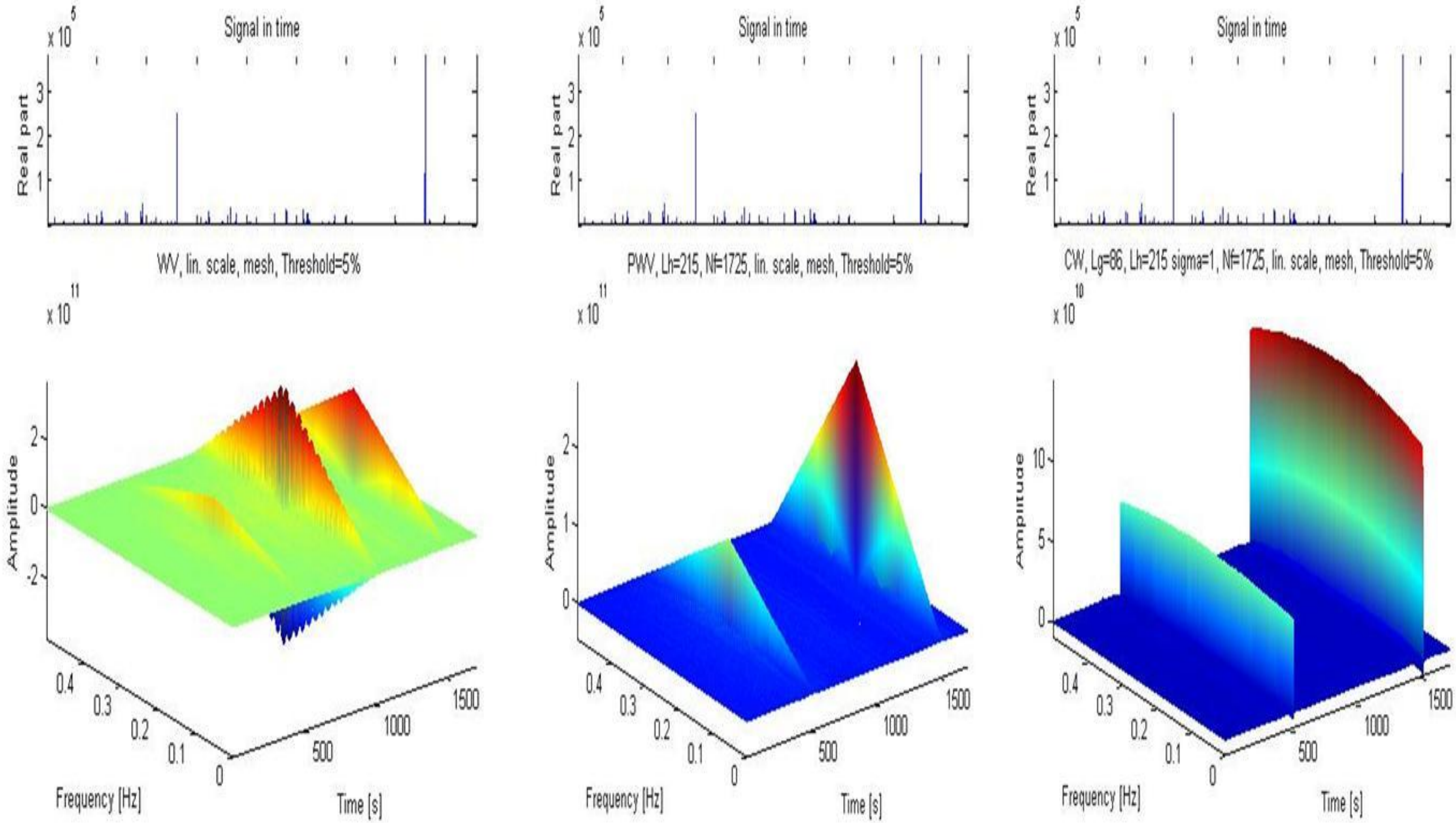
- Extensive port and host-behaviour-based approach
- Usage of *graphlets* from BLINC



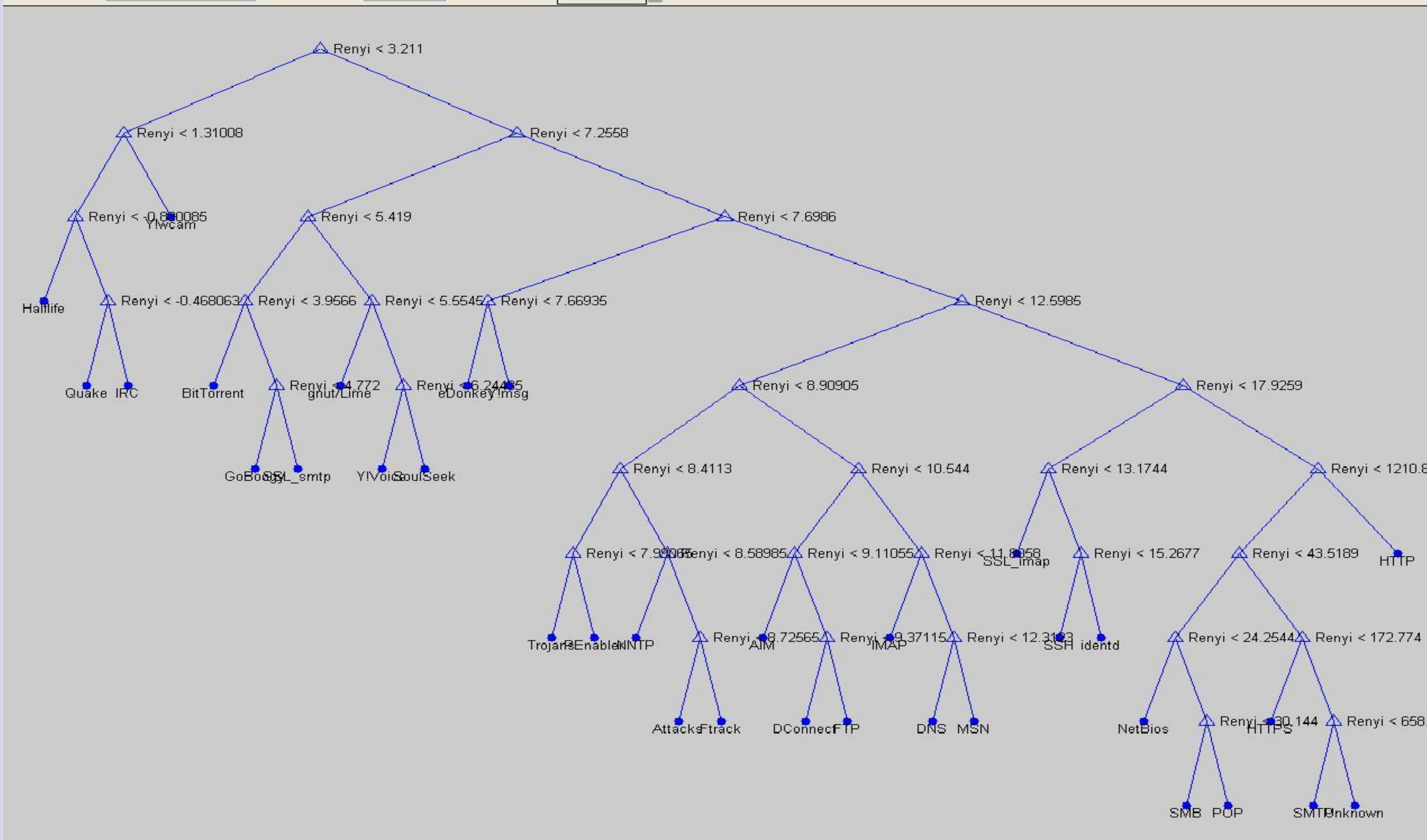
Pre-processing for Traffic Classification (cont..)

- Keio-I : training set , Keio-II : test set
- Computation of each energy distribution for every application protocol individually based on the packet and byte-wise utilization of TCP & UDP.
- Comparison between distributions.
- Extraction of the Renyi Dimension for every application protocol from the selected TF distribution.

Comparison of energy TF distributions (example : Keio TCP bytes for MSN)



Results (example: Classification of TCP bytes for Keio trace - SPWV)



Results (cont)

- Overall Accuracy

Keio trace : 95%(pkts)
93%(bytes)

WIDE trace : 92% (pkts)
88% (bytes)

Traffic Cat.	Recall% (bytes)	Recall% (pkts)
WWW	>=90.4%	>=95.8%
FTP	>=94.5%	>=97.3%
P2P	>=84.8%	>=91.9%
DNS	>=95.6%	>=98.6%
Mail/News	>=93.3%	>=97.8%
Streaming	>=81.3%	>=92.2%
Net. Ops.	>=96.8%	>=94.1%
Encryption	>=95.3%	>=89.8%
Games	>=89.3%	>=93.9%
Chat	>=82.1%	>=92.7%
Attack	>=78.9%	>=88.6%

Summary

- Backbone and Edge network link traffic is highly non-stationary.
- Suitability of Energy TF distributions for general traffic profiling.
- Practical usability presented particularly in the area of traffic classification.
- Introduction of complexity-based traffic classification based on the 3rd order Renyi Dimension.
- Packet-based analysis indicated higher accuracy.

On going &Future Work

- New network-oriented features (e.g. 5 tuple)
- New Energy TF metrics (e.g. 1st , 2nd order moment sequence)
- Employment of Support Vector Machines.
- Full, comparison with BLINC on larger datasets.

Thank you ☺