State of the Art in Traffic Classification:
A Research Review

min zhang
wolfgang john
kc claffy
nevil browlee

Outline

- Motivation
- Research review and taxonomy
- Survey analysis: P2P
- Discussion and conclusion

Motivation

- Today's Internet
  - evolving in scope and complexity
  - applications adapt rapidly to detection attempts
  - emerging obfuscation techniques
- Many classification approaches in literature
  - using whatever traffic samples available
  - no systematic integration of results

Motivation contd.

- Filling this gap, our research review
  - creates a structured taxonomy of traffic classification papers and their datasets
  - helps to answer popular questions
  - reveals open issues and challenges

Research review and taxonomy

- 64 papers published between 1994 and 2008

- Definition: *traffic classification*

  Methods of classifying traffic **data sets** based on **features** passively observed in the traffic, according to specific **classification goals.**

http://www.caida.org/research/traffic-analysis/classification-overview

Research review and taxonomy contd.

- Data sets: more than 80 data sets used for 64 papers!

  Categorized by: Time of collection, link type, capture environments, geographic location, payload length, etc

- Classification goals: coarse or finer-grained

Research review and taxonomy contd.

- Features

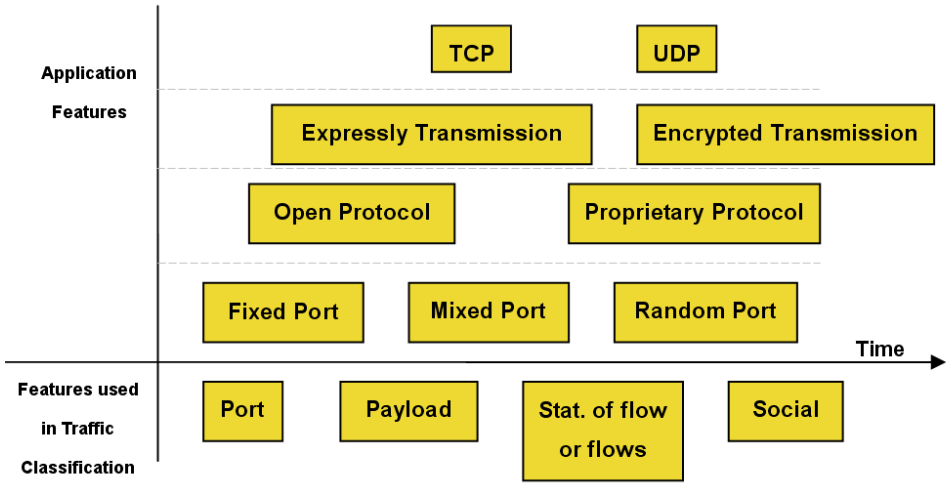

Figure 1: Trends of applications and features

Research review and taxonomy contd.

- Methods
  - exact matching: port number, payload, etc
  - heuristic methods, e.g. on connection patterns
  - machine learning methods:

    supervised and unsupervised

http://www.caida.org/research/traffic-analysis/classification-overview

Survey analysis: P2P

- How much P2P?

  1.2% to 93% across the 18 (out of 64) papers

Table 1: P2P Percentage of Year

Year	Range of P2P Volume	Paper
2002	21.5%	[14]
2004	9.19-60%	[9],[10],[11],[6],[16]
2006	35.1-93%	[3],[5],[4],[8]

Table 2: P2P Percentage of Link Location

Year	Link Location	Range of P2P Volume	Paper
2004	Campus link	31.3%	[11]
2004	ADSL link	60%	[16]
2004	Backbone link	9-14%	[9],[6]
		17-25%	[10]

Survey analysis: P2P contd.

- How much P2P? (cont')

Table 3: P2P Percentage of Geographic Location

Geo Location	Year	Range of P2P Volume	Paper
Europe	2005	60-80%	[15]
	2006	79-93%	[7],[8]
North America	2003	8%,10.7%	[9]
	2004	14%, 9.9%	[9]
	2003-04	9.2-70%	[10],[6],[12]
	2006	21-35%	[3],[5],[4]
Asia	2002	21.5%	[14]
	2005	1.34% (port-based)	[2]
	2008	1.29% (port-based)	[2]

Discussion and Conclusions

- Shortcomings of current traffic classification efforts:
  - 80 data sets by 64 papers → lack of shared, current data sets as reference data
  - no clear definition of P2P or file-sharing → lack of standardized measures and classification goals
  - Poor comparability of results!!!

Discussion and Conclusions contd.

- So how much of modern Internet traffic is P2P?

  "there is a wide range of P2P traffic on Internet links; see your specific link of interest and classification technique you trust for more details."

- This review can answer further questions:
  - TCP/UDP ratio?
  - Amount of encrypted traffic?
  - Tunneled traffic?
  - …

- Thanks
- 謝謝
- Bedankt
- Merci
- Danke
- Ε υ χ α ρ ι σ τ ώ
- Grazie
- あ り が と う
- 감사합니다
- Dzieki
- Gracias
- شكرا

http://www.caida.org/research/traffic-analysis/classification-overview/