State of the Art in Traffic Classification: A Research Review

Min Zhang
Beijing Jiaotong University
mia.minzhang@gmail.com

Wolfgang John
Chalmers University, Sweden
wolfgang.john@chalmers.se

KC Claffy, Nevil Brownlee
CAIDA, UC San Diego
{kc,nevil}@caida.org

1. INTRODUCTION

The Internet, while emerging as the key component for all sorts of communication, is far from well-understood. The goal of traffic classification is to understand the type of traffic carried on the Internet, which continually evolves in scope and complexity. For security and privacy reasons, many applications have emerged that utilize obfuscation techniques such as random ports, encrypted data transmission, or proprietary communication protocols. Further, applications adapt rapidly in the face of attempts to detect certain types of traffic, creating a challenge for traffic classification schemes. Research papers on Internet traffic classification try to classify whatever traffic samples a researcher can find, with no systematic integration of results. With the exception of machine learning techniques for traffic classification[13], we know of no complete overview of traffic classification attempts. To fill this gap, we have created a structured taxonomy of traffic classification papers and their datasets. To illustrate its utility, we use the taxonomy to answer the recently most popular question about traffic (*"How much is peer-to-peer file sharing?"*). Our survey also reveals open issues and challenges in traffic classification.

2. RESEARCH REVIEW

Our review is based on 64 papers published between 1994 and 2008, starting with papers from top-ranked, peer-reviewed academic research conferences, and then including papers cited from this seeding set of papers, as well as follow-up papers written by the same authors.

We use the phrase *traffic classification* to refer to **methods** of classifying traffic **data sets** based on **features** passively observed in the traffic, according to specific **classification goals**. On a supplementary web page [1], we group papers into five categories: *survey, analysis, methodology, tools* and *others*. Analysis papers seek trustworthy numbers on traffic composition, while methodology papers focus on the methods of classification. We also provide a flexible, interactive table that supports selection of relevant attributes of papers, e.g., data sets, methods, goals, main findings, etc.

2.1 Data Sets

Several public and private passive measurement infrastructures have provided a variety of data sets for Internet traffic classification studies. Based on our analysis, we find that these 64 papers make use of more than 80 data sets, which we classify based on *time of collection, link type, capture environments, geographic location, payload length, etc.*

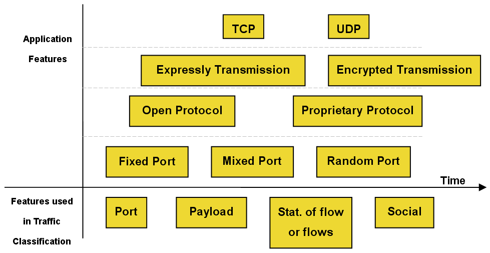

Figure 1: Trends of applications and features

2.2 Classification Goals and Features

Although traffic classification is a rather specific research field, the goals of these research papers are not identical. Some only have *coarse classification* goals, i.e., whether it's transaction-oriented, bulk-transfer, or peer-to-peer file sharing. Some have a *finer-grained classification* goal, i.e., the exact application generating the traffic.

Selection of traffic features used for classification evolves with application development. Media-rich entertainment applications - and associated attempts to discriminate against such applications - have inspired sophisticated obfuscation methods. Fig.1 gives a rough view of application and classification features. Recently, some applications (uTorrent, PPStream, PPLive) have changed from using TCP to UDP, a dramatic challenge for traditional traffic engineering.

Fifteen years ago, researchers could reasonably accurately classify traffic using TCP or UDP *port numbers*, but as applications began to use unpredictable ports, accurate classification requires *payload* examination. Examining payload is a controversial methodology due to privacy concerns, and is not even possible for encrypted payload, so researchers have also studied techniques that are independent of packet content, such as *statistical features* based on network flows or underlying *social networks* to identify per-host behavior.

2.3 Methods

Methods to classify traffic at an application level include *exact matching*, e.g., of port number and payload; *heuristic methods*, applied e.g. on connection patterns to infer social networks; or *machine learning* based on statistical features. We group machine learning methods into two categories: *Supervised Learning* and *Unsupervised Learning*. Naive Bayes, Decision Tree, NN, LDA, QDA, Bayesian Neural network are supervised learning algorithms; EM, AutoClass and K-Means are unsupervised learning algorithms [13].

PAM2009, April 1-3, 2009, Seoul, Korea.
.

3. SURVEY ANALYSIS: HOW MUCH P2P?

P2P traffic is one of the most challenging traffic types to classify. This is the result of substantial legal interest in identifying it and even more substantial negative repercussions to the user if P2P traffic is accurately identified. The misaligned incentives between those who want to use and those who want to identify P2P applications, together with the tremendous legal and privacy constraints against traffic research, render scientific study of this question near impossible. Even if possible, wide variation across links would prevent a simple numeric answer to the question of how much P2P traffic there is on the Internet.

Nonetheless, our taxonomy does reveal insights: the fraction of peer-to-peer file sharing traffic observed ranges from 1.2% to 93% across the 18 (out of 64) papers that provide such numbers. We also know that the average fractions reported have increased considerably from 2002 to 2006 (Table 1). Tables 2 and 3 show that results also vary widely by link and geographic location. Table 3 suggests that P2P is more popular in Europe, probably due to stricter policies (MPAA and RIAA) in North America. Note that the Asian results are from Japanese data sets, in which 1.34% and 1.29% are based on port numbers and therefore likely to significantly underestimate the fraction of P2P traffic. Furthermore, the amount of P2P traffic also varies by time of day, with higher fractions at night [5, 8].

One study[5] suggests that peer-to-peer applications are used more often at home than in the office. Finally, a study[8] in Europe found a higher fraction of P2P traffic on an European university link than some Canadian academics[5] found on their campus. Many of these numbers are based on statistical or host-behvioral classification, not the most reliable methods of detecting applications. More accurate methods involve examination of traffic contents (if unencrypted), which is fraught with legal and privacy issues.

Our taxonomy can allow similar analyses of other open questions, such as trends and development of traffic classes or features, yielding new insights into Internet traffic.

Table 1: P2P Range (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 Range (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]

Table 3: P2P Range (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]

4. DISCUSSION

This research review, including 64 papers and more than 80 data sets, shows that traffic classification methods have evolved in response to the more sophisticated obfuscation techniques of network applications. We present a rough taxonomy of traffic classification approaches, based on features, methods, goals and data sets.

Our survey review also reveals shortcomings with current traffic classification efforts. First of all, the variety of data sets used does not allow systematic comparison of methods. Few research groups (can) share their datasets. Already true ten years ago, the field of traffic classification research still needs publicly available, modern data sets as reference data for validating approaches. This need however requires clear policies for data sharing, including accepted anonymization and desensitization guidelines. Secondly, the lack of standardized measures and classification goals is further amplifying the poor comparability of results. For example, there exists no clear definition for traffic classes such as P2P or file-sharing.

Despite these shortcomings, we showed how the taxonomy can shed insight on questions such as: *”how much of modern Internet traffic is P2P?”* Though we found some trends and indications, we have far too little data available to make conclusive claims beyond *”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.”*

5. REFERENCES

[1] CAIDA. An overview of traffic classification, 2009. *http://www.caida.org/research/traffic-analysis/classification-overview/.*

[2] K. Cho, K. Fukuda, H. Esaki, and A. Kato. Observing slow crustal movement in residential user traffic. *CoNEXT*, 2008.

[3] J. Erman, M. Arlitt, and A. Mahanti. Traffic classification using clustering algorithms. *SIGCOMM*, 2006.

[4] J. Erman, A. Manhanti, M. Arlitt, I. Cohen, and C. Williamson. Identifying and discrimination between web and peer-to-peer traffic in the network core. *WWW*, 2007.

[5] J. Erman, A. Manhanti, M. Arlitt, I. Cohen, and C. Williamson. Offline/realtime traffic classification using semi-supervised learning. *Perform. Eval*, 2007.

[6] M. Iliofotou, P. Pappu, and M. Faloutsos. Graption: Automated detection of p2p applications using traffic dispersion graphs. *Technical Report*, 2008.

[7] W. John and S. Tafvelin. Heuristics to classify internet backbone traffic based on connection patterns. *ICOIN*, 2008.

[8] W. John, S. Tafvelin, and T. Olovsson. Trends and differences in connections behavior within classes of internet backbone traffic. *PAM*, 2008.

[9] T. Karagiannis, A. Broido, N. Brownlee, K. Claffy, and M. Faloutsos. Is p2p dying or just hiding? *GLOBECOM*, 2004.

[10] T. Karagiannis, A. Broido, M. Faloutsos, and K. Claffy. Transport layer identification of p2p traffic. *SIGCOMM*, 2004.

[11] T. Karagiannis, K. Papagiannaki, and M. Faloutsos. Blinc multilevel traffic classification in the dark. *SIGCOMM*, 2005.

[12] A. Madhukar and C. Williamson. A longitudinal study of p2p traffic classification. *MASCOTS*, 2006.

[13] T. Nguyen and G. Armitage. A survey of techniques for internet traffic classification using machine learning. *IEEE Communications Surveys*, 2008.

[14] M. Perenyi, T. Dang, A. Gefferth, and S. Monlnar. Flow analysis of internet traffic: World wide web versus peer-to-peer. *System and Computers in Janpan*, 2005.

[15] M. Perenyi, T. Dang, A. Gefferth, and S. Monlnar. Identification and analysis of peer-to-peer traffic. *Journal of Communications*, 2006.

[16] L. Plissonneau, J. Costeux, and P. Brown. Analysis of peer-to-peer traffic on adsl. *PAM*, 2005.