ISSN 2350 - 0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 6, June 2015

An In-depth Understanding Of Internet Backbone Traffic : A Case Study Approach

Madhusmita Panda

Department of Computer Science and Engineering, C V Raman College of Engineering, BPUT, Bhubaneswar, Odisha, India

ABSTRACT: Internet is very essential in day to day life of a common citizen. Modeling the Internet traffic is an important issue. Internet Protocol network capacity planning is a very important task. The Internet backbone is the principal data routes between large, strategically interconnected computer networks and core routers on the Internet. These data routes are hosted by commercial, government, academic and other high capacity network centres, the Internet exchange points and network access points that interchange Internet traffic between the countries, continents and across the oceans. Internet service providers, often Tier 1 networks, participate in Internet backbone exchange traffic by privately negotiated interconnection agreements, primarily governed by the principle of settlement free peering. We study the understanding of Internet traffic characteristics by measuring and analyzing modern Internet backbone data. We start the thesis with an overview of several important considerations for passive Internet traffic collection on large-scale network links. The lessons learned from a successful measurement project on academic Internet backbone links can serve as guidelines to others setting up and performing similar measurements. The data from these measurements are the basis for the analyses made in this thesis. As a first result we present a detailed characterization of packet headers. We propose a method and accompanying metrics to assess routing symmetry on a flow-level based on passive measurements. This method helps to improve traffic analysis techniques. TCP as the transport protocol in backbone traffic. We observe an increase of UDP traffic during the last few years, which we attribute to P2P signalling traffic. These results show the detailed classification of traffic according to network application. To accomplish this, we review state-of-the-art traffic classification approaches and subsequently propose two new methods. The first method provides a payload-independent classification of aggregated traffic based on connection patterns, second, classification method for fine-grained protocol identification by utilizing statistical packet and flow features. This method is capable of accurate classification in a simple and efficient way. This thesis presents methods and results contributing additional perspectives on Internet classifications and characteristics at different levels of granularity.

KEYWORDS: Internet, IP, UDP, TCP, P2P, traffic analysis.

I. INTRODUCTION

Today, the Internet has emerged as the key component in personal and commercial communication. One contributing factor to the ongoing expansion of the Internet is its versatility and flexibility. In fact, almost any electronic device can be connected to the Internet these days, ranging from traditional desktop computers, servers and supercomputers to all kinds of wireless devices, embedded systems, sensors and even home equipment. Accordingly, the usage of the Internet has changed dramatically since its initial operation in the early 1980s, when it was a research project connecting a handful of computers, facilitating a small set of remote operations. Today the Internet serves as the data backbone for all kinds of protocols, making it possible to interact and exchange not only text, but also voice, audio, video, and various other forms of digital media between hundreds of millions of nodes.

Traditionally, an illustration of the protocol layers of the Internet pictures an hourglass, with a single Internet Protocol (IP) on the central network layer and an increasingly broader spectrum of protocols above and below. Since the introduction of IP in 1981, a protocol that is basically still unchanged, technology and protocols have developed significantly. Underlying transmission media evolved from copper to fiber optics and wireless technologies, routers and switches became more intelligent, and are now able to handle Gbit/s instead of Kbit/s, and additional middleware devices have been introduced (e.g., Network Address Translation boxes and firewalls). Above the network layer, new applications have also constantly been added, ranging from basic services such as the Domain Name System (DNS) and Hypertext Transfer Protocol (HTTP), to recent complex peer-to-peer (P2P) protocols allowing applications for file sharing, video streaming, and IP telephony. With the introduction of IPv6, even the foundation of the Internet, IP, is finally about to be substituted.

This multiplicity of protocols and technologies leads to a continuous increase in the complexity of the Internet as a whole. Of course, individual protocols and network infrastructures are usually well understood and tested in isolated lab environments or simulations. However, their behaviour as observed while interacting with the vast diversity of applications and technologies in the Internet environment is often unclear, especially on a global scale.

This lack of understanding is further amplified by the fact that the topology of the Internet was not planned in advance. The current Internet topology is the result of an uncoordinated extension process, where heterogeneous networks of independent organizations have been connected one by one to the main Internet (INTERconnected NETworks). As a consequence, the Internet today is built up of independent, autonomous network systems, where each autonomous system (AS) has its own set of usage and pricing policies, quality of service (QoS) measures and resulting traffic mix. Thus, the usage of Internet protocols and applications is not only changing over time but also with geographical locations .Finally, higher connectivity bandwidths, growing numbers of users and increasing economical importance of the Internet also lead to an increase in misuse and anomalous behaviour .Not only do the numbers of malicious incidents continue to rise, but also the level of sophistication of attack methods and available tools. Today, automated attack tools employ advanced attack patterns and react on the deployment of firewalls and intrusion detection systems by cleverly obfuscating their malicious actions. Malicious activities range from host- and port-scanning to more sophisticated attack types, such as worms and various denials of service attacks. Unfortunately, the Internet, initially meant to be a friendly place, eventually became a very hostile environment that needs to be studied continuously in order to develop suitable counter strategies.

For the reasons mentioned above, network researchers and engineers currently have limited understanding of the modern Internet, despite its emergence as a critical infrastructure of global importance. We identified a number of important open questions that Internet measurements help to answer. We grouped them into four rough categories:

(i) *Scalability and sustainability* issues regarding fundamental Internet services, including routing scalability, AS level topology evolution, IP address space utilization, DNS scalability and security;

(ii) *Internet performance*, e.g., the impact of new protocols and applications on Internet performance characteristics such as per-flow throughput, jitter, latency and packet loss/reordering;

(iii) *Evolution of Internet traffic*, such as traffic growth trends, protocol and application mix at different times and different locations;

(iv) *Network security*, including anomaly detection and mitigation of network attacks and other unwanted/unsolicited traffic, such as email spam, botnet and scanning traffic. Given the usability to collect Internet traffic data on a wide-area network backbone link, this thesis addresses the tter two categories. We also discuss methodological aspects of passive Internet measurement and data collection, which form the basis for our results. Specifically, the thesis sets out to provide a better understanding of the modern Internet by presenting current characteristics of Internet traffic ased on a large amount of empirical data. We claim that it is crucial for the Internet community to understand the nature and detailed behaviour of modern network traffic. A deeper understanding would support optimization and development of network protocols and devices, and further improve the security of network applications and the protection of Internet users.

Before we could perform offline analysis of Internet traffic, we had to set up a measurement infrastructure to collect Internet data. Packet-level data collection on large-scale net-work links, however, is a non-trivial task. One reason for the difficulties is the rapid and decentralized development of the Internet on a competitive market, which historically has left both little time and few resources to integrate measurement and analysis possibilities into Internet infrastructure, applications and protocols. Traditional network management therefore relies on aggregated measurements from individual nodes, such as SNMP Statistics (Simple Network Management Protocol and statistics of sampled flow data (e.g., flow counts and flow throughputs, size and duration distributions). However, we claim that we in addition need complete, fine grained data in order to obtain a comprehensive and detailed understanding of the modern Internet. Empirically measured Internet datasets constitute an important data source for different purposes:

Active vs. passive measurement approaches: Active measurement involves injecting traffic into the network to probe certain network devices (e.g., ping) or to measure network properties such as Round Trip Times (RTT), one-way delay and maximum bandwidth. Passive measurement or monitoring based on pure observation of network traffic is non-intrusive and does not change the existing traffic. Network traffic is tapped at a specific location and can then be recorded and processed at different levels of granularity, from complete packet-level traces to only a statistical summary.

II. LITERATURE SURVEY

We define traffic characterization as the analysis of Internet data resulting in a description of traffic properties. These properties can range from the features of aggregate network traffic to detailed features of single packets and flows. Specifically, traffic characterization in this thesis covers a detailed, fine grained traffic analysis of packet and flow-level data. Since the Internet is a moving and continually evolving target, some of our results revise or update previous studies that are based on outdated data sets collected years ago. Our packet-level characterization reveals general traffic properties such as packet size distribution and transport protocol breakdown and also shows the current deployment of protocol-specific features such as IP and TCP options and flags , which is relevant input to Internet simulation models. Our packet analysis furthermore includes a systematic listing of packet header anomalies together with their frequencies on the observed Internet backbone links, which provides an empirical background for the development and refinement of traffic filters, firewalls and intrusion detection systems. Furthermore, we believe that knowledge of such detailed Internet traffic characteristics can help researchers and practitioners in designing networked devices and applications and in improving their performance and robustness. Flow-level analysis aggregates individual packets into flows, which can provide additional insights into traffic characteristics. We propose a method and accompanying metrics to assess routing symmetry flow measurements from a specific link, and the results suggest that routing symmetry is uncommon on non-edge Internet links. We then confirm the predominance of TCP as transport protocol in backbone traffic, but note an increase of UDP traffic during the last few years. These results verify common assumptions about Internet traffic, which are often, embedded into traffic analysis or classification tools . Consequently, the results can impact advanced Internet analysis efforts and provide further measurement support for Internet modelling. We also provide a detailed analysis of TCP flows to reveal network properties such as connection lifetime, size and establishment/termination behaviour. The results of this flow analysis highlight the need for traffic classification according to application as a next step towards a better understanding of Internet traffic behaviour. The following analysis of classified traffic reveals trends and differences in connection properties of Internet traffic and shows how different classes are there. These results show how current transport protocols are utilized by application developers, facilitating the improved design of network devices, software, and protocols. In this thesis we apply a passive measurement approach to provide analysis of Internet backbone traffic properties.

We define traffic classification as the analysis of Internet data resulting in a decomposition of the traffic according to application layer protocols Traffic classification results can be useful for traffic management purposes and traffic engineering purposes such as optimization of network design and resource provisioning. Network applications have been designed to use well known port numbers to communicate with servers or peers, making traffic classification relatively straightforward. However the developers of upcoming file sharing applications started to deviate from the standard behaviour by using dynamic port numbers, thus diminishing the accuracy of port-based classification. Since then, there has been an ongoing arms race between application developers trying to avoid traffic filtering or classification, and operators, network researchers, and other institutions interested in accurate traffic classification. Researchers first used static payload examination to classify applications using unpredictable ports, an approach also used in commercial tools. Application developers then reacted by using proprietary protocols and payload encryption, which means that modern traffic classification methods cannot rely solely on port number information and static payload signatures .

Internet traffic is the flow of data across the Internet. Because of the distributed nature of the Internet, there is no single point of measurement for total Internet traffic. Internet traffic data from public peering points can give an indication of

Internet volume and growth, but these figures exclude traffic that remains within a single service provider's network as well as traffic that crosses private peering points. A quick way to understand the volume of traffic in some part of the Internet is to read the current latency figures that are being reported on The Internet.

The packet counts with the procees are Poisson process and Self similar process

A Poisson process

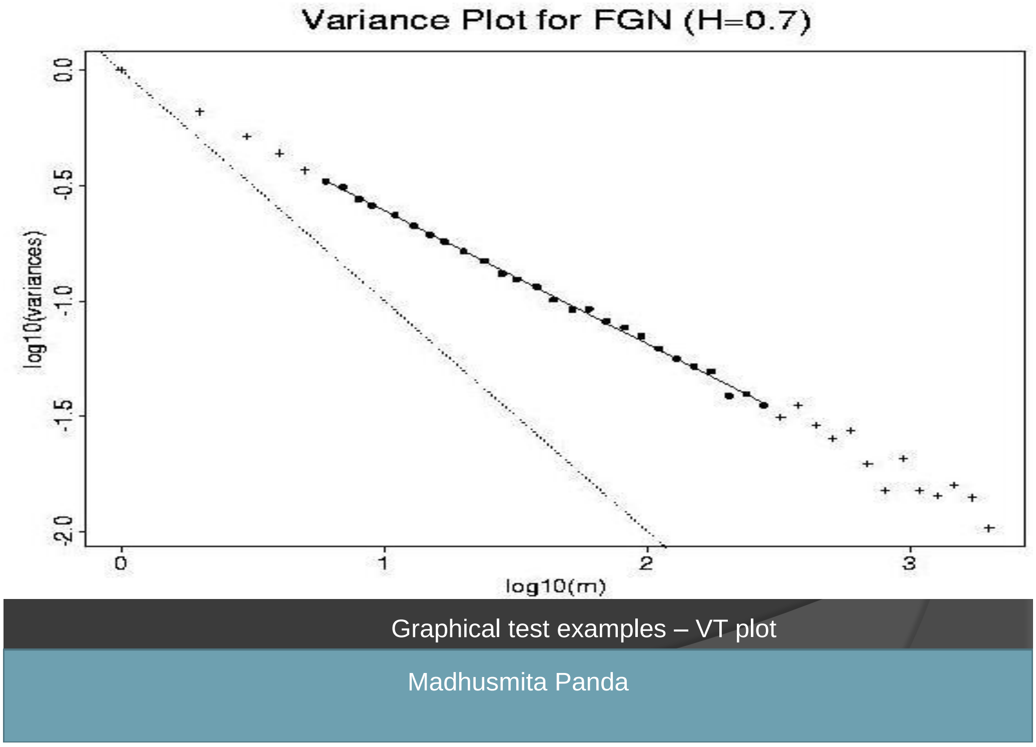
- When observed on a fine time scale will appear bursty
- When aggregated on a coarse time scale will flatten (smooth) to white noise

A Self-Similar (fractal) process

- When aggregated over wide range of time scales will maintain its bursty characteristic

Graphical Tests for Self-Similarity

- Variance-time plots
  - Relies on slowly decaying variance of self-similar series
  - The variance of X$^{(m)}$ is plotted versus *m* on log-log plot
  - Slope (-b) greater than –1 is indicative of SS
- R/S plots
  - Relies on rescaled range (R/S) statistic growing like a power law with H as a function of number of points *n* plotted.
  - The plot of R/S versus *n* on log-log has slope which estimates H
- Periodogram plot
  - Relies on the slope of the power spectrum of the series as frequency approaches zero
  - The periodogram slope is a straight line with slope b – 1 close to the origin



Graphical test examples – VT plot

Madhusmita Panda

Figure 2.1 Graphical test examples – VT plot

III. CHARACTERIZATION OF INTERNET TRAFFIC

Packet-level Characterization

Analysis of Internet Backbone Traffic and Anomalies Observed reflects packet characteristics on SUNET Internet backbone traffic and points out misbehaviours and potential problems. We used the bidirectional traffic collected on GigaSUNET.

Results

The 148 traces analysed sum up to 10.77 billion PoS frames, containing a total of 7.6 TB of data. 99.97% of the frames contain IPv4 packets, summing up to 99.99% of the carried data. The remaining traffic consists of different routing protocols (BGP, CLNP, CDP). The results in the remainder of this paper are based on IPv4 traffic only.

Transport protocols

The protocol breakdown in Table 3.1(a) once more confirms the dominance of TCP traffic. Compared to earlier measurements reporting about TCP accounting for around 90 - 95% of the data volume and for around 85-90% of IP packets, both fractions seem to be slightly larger in the analysed SUNET data. In Table 3.1(a), the fractions of cumulated packets and bytes carried in the respective protocol are given in percent of the total IPv4 traffic for the corresponding time. An interesting observation can be made at the 2PM data. Here, the largest fraction of TCP and the lowest of UDP packets appear. A closer look at the differences between outgoing and incoming traffic revealed that three consecutive measurements on the outgoing link carried up to58% UDP packets, not covering the 2PM traces, as shown in Table 3.1 (b). These figures indicate a potential UDP burst of 14-24 hours of time. A detailed analysis showed that the packet length for the UDP packets causing the burst was just 29 bytes, leaving a single byte for UDP payload data. These packets were transmitted between a single sender and receiver address with varying port numbers. After reporting this network anomaly, the network support group of a University confirmed that the burst stemmed from an UDP DoS script installed undetected on a web server with a known vulnerability. Although TCP data was still predominant, a dominance of UDP packets over such a time span could potentially lead to TCP starvation and raise serious concerns about Internet stability and fairness.

	2AM		10AM		2PM		8PM	
	Pkts	Data	Pkts	Data	Pkts	Data	Pkts	Data
TCP	91.3	97.6	91.5	96.8	93.2	97.1	91.4	97.2
UDP	8.5	2.3	7.6	2.8	6.1	2.7	8.3	2.7

(a) IPv4 Protocol Breakdown (values in %)

Analysis of IP Properties

The TOS field can optionally include code points for Explicit Congestion Notification (ECN) and Differentiated Services. 83.1% of the observed IPv4 packets store a value of zero in the TOS field, not applying the mechanisms above. Valid ’Pool 1’ DiffServ Codepoints (RFC 2474) account for 16.8% of all TOS fields. Medina et al. Reported about almost a doubling of ECN capable web servers from 1.1% in 2000 to 2.1% in 2004, but indicates that routers or middle boxes might erase ECT code points. In our data only 1.0 million IPv4 packets provide ECN capable transport (either one of the ECT bits set) and additionally 1.1 million packets actually show 'congestion experienced' (both bits set). This means that ECN is implemented in only around 0.02% of the IPv4 traffic. These numbers are consistent with the observations by Pentikousis et al., suggesting that the number of ECN-aware routers is still very small.

CLASSIFICATION OF INTERNET TRAFFIC

Traffic classification can be applied for various purposes such as traffic management, traffic engineering, security monitoring, accounting, policy enforcement that require different classification granularity

Flow-based Symmetry Estimator

In this section we present the Flow-based Symmetry Estimator (FSE), a simple method depicted below and associated tool to estimate the level of routing symmetry from passively measured flow data that takes unidirectional 5-tuple flow data as input. We could have computed symmetry based on IP pairs (2-tuples), but most traffic classification and engineering methods deal with flows, so we chose the flow granularity. Due to its simplicity, most traffic analysis tools prefer this method to tracking TCP connection state, although we use TCP connection information extracted from packet level data to validate our technique.

The FSE method

1: given a time-interval of traffic trace:

2: consider TCP data traffic (TCP packets carrying data)

3: Tf (Tb) = set of tuples going forward (backward)

4: Tf\ Tb = set of symmetric tuples TS

5: pkts (bytes) in TS=set of symmetric pkts (bytes)

After collecting a unique list of unidirectional flows for each direction of a link, FSE classifies 5-tuples as symmetric if they appear on both lists. Packet (byte)-level symmetry is the fraction of packets (bytes) sent between tuples classified as symmetric, so that the degree of symmetry can be quantified in three dimensions: 5-tuple flows, packets, bytes.

IV. RESULTS AND DISCUSSIONS

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.

One study suggests that peer-to-peer applications are used more often at home than in the office. Finally, a study in Europe found a higher fraction of P2P traffic on a European university link than some Canadian academics found on their campus. Many of these numbers are based on statistical or host-behavioural 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.

Year	Range of P2P Volume
2002	21.5%
2004	9.19-60%
2006	35.1-93%

Table 4.1: P2P Range (Year)

Year	Link Location	Range of P2P Volume
2004	Campus link	31.3%
2004	ADSL link	60%
2004	Backbone link	9-14%

Table 4.2: P2P Range (Link Location)

Results and discussion

We applied the proposed heuristics to our data traces represents time series of classified network protocols. The x-axis of the graphs represents time, with one bar for each trace time. Four traces on three days (07/04, 09/04, 23/04) had to be discarded due to measurement errors. The remaining whitespaces between bars represent the 8 hour measurement break between 2AM and 10AM, which means that each continuous block represents 4 traces collected in the order of different times. In the connection breakdown, only four categories are visible, since flows classified by H5 are too small in number to show up in this graph. Anyhow, these 31,000 long flows are responsible for almost 10% of the TCP data. Typically, these flows begin and end outside the measurement period and transfer data between hosts, which do not generate additional traffic on our links. Since our classification method is based on connection patterns, insufficient connection numbers for a particular host reveal a weakness of this method. In the data breakdown on the other hand, flows classified by F9 (attacks) are not visible. Even though attacks represent between 8 and 60% of the flows, they carry less than 1% of the data on average. This also proves the power of F9, since it effectively detects DoS attacks and network scanning, which typically show up as short 1-packet flows only, carrying no payload data. P2P flows (flows matching H1-H4, while not matching any of the false positive rules F1-F10) account for an average of 42% of the connections. On the other hand, they carry between 66 and 87% of the traffic, with an average of 79%. This indicates once more the success of the heuristics, since P2P flows are expected to carry more data on average than non-P2P flows. On this dataset, the proposed heuristics left as little as 1% of the connections and 0.2% of the data unclassified (except the flows classified by H5). While a careful analysis of these results need to be done as future work.

V.CONCLUSION

The role of the Internet as a critical infrastructure of global importance. Optimization and development of network protocols and devices, and furthermore helps to improve the security of network applications and the protection of Internet users. In this thesis, we therefore presented methods and results contributing additional perspectives on global

Internet behaviour at different levels of granularity. We study the methodological aspects of passive Internet measurement and data collection.

The results, representing actual, empirically measured properties of network traffic, are important for scientific network simulation and modeling and are relevant as well for operational purposes such as network management, traffic engineering, and network security.As indicated in the study of UDP traffic, assumptions that have been valid for a long time can be misleading if they are not revisited periodically and from varying points. In the MonNet project, we made most of the analyses with in-house tools developed from scratch. important aspect of standard tools and APIs is the possibility to easily compare results with those from related tools. Our results indicate that exploitation of statistical features is a promising method for reliable traffic classification. This approach may prove useful especially in the face of non-existing or obfuscated payload and further complicating circumstances such as unidirectional traffic flows and large fractions of UDP traffic, but it also requires at least partial access to privacy sensitive packet payload. Traffic classification can be applied for various purposes such as traffic management, traffic engineering, security monitoring, accounting, policy enforcement that require different classification granularity. We find that the lack of singularly defined traffic classes further amplifies the poor comparability of classification results. To give an example, for some purposes, Skype traffic might be regarded as P2P traffic, but not P2P file sharing, while for other purposes it could be classified as voice-over-IP.We believe that the research community would benefit from a set of common definitions for traffic classes in different granularities depending on the purpose. We conclude that measuring traffic on large-scale Internet links is a tedious task, which can be both very expensive and time consuming. However, Internet measurement research, empirical in nature, depends on the quality and diversity of available network traces. We therefore identify a further challenge: the complications of sharing or providing access to the tediously collected data, which is related to the legal and ethical limbo of scientific Internet data collection. Lack of available datasets is a major shortcoming of current traffic classification efforts as well as any other type of traffic analysis To increase the credibility of Internet measurement as a research discipline, we therefore argue that it is essential for the research community to agree on ways to facilitate the sharing of network data in a manner that balances the privacy requirements of data owners and providers with the information requirements of researchers (i.e., ways to handle the trade-off between data privacy and utility).

REFERENCES

[1] M. Zhang, W. John, kc claffy, and N. Brownlee, “State of the Art in Traffic Classification: A Research Review,” *PAM Student Workshop*, 2009.

[2] A. W. Moore and K. Papagiannaki, “Toward the Accurate Identification of Network Applications,” *PAM: Passive and Active Measurement Conference*, 2005.

[3] A. Madhukar and C.Williamson, “A Longitudinal Study of P2P Traffic Classification,” *MASCOTS*, 2006.

[4] S. Sen, O. Spatscheck, and D. Wang, “Accurate, Scalable In-network Identification of P2P Traffic Using Application Signatures,” *WWW*, 2004.

[5] L7-filter, “Application layer packet classifier for linux,” 2009, http://l7-filter.sourceforge.net/ (accessed 2009-04-02).

[6] T. Karagiannis, K. Papagiannaki, and M. Faloutsos, “BLINC Multilevel Traffic Classification in the Dark,” *SIGCOMM*, 2005.

[7] W. John and S. Tafvelin, “Heuristics to Classify Internet Backbone Traffic based on Connection Patterns,” *ICOIN*, 2008.

[8] M. Iliofotou, H. Kim, M. Faloutsos, M. Mitzenmacher, P. Pappu, and G. Varghese, “Graph-based P2P Traffic Classifcation at the Internet Backbone,” *IEEE Global Internet Symposium*, 2009.

[9] J. Erman, M. Arlitt, and A. Mahanti, “Traffic Classification Using Clustering Algorithms,” *SIGCOMM*, 2006.

[10] M. Crotti, M. Dusi, F. Gringoli, and L. Salgarelli, “Traffic Classification Through Simple Statistical Fingerprinting,” *SIGCOMM Comput. Commun. Rev.*, vol. 37, no. 1, pp. 5–16, 2007.

[11]E.Hjelmvik, “The SPID Algorithm - Statistical Protocol Identification,” www.iis.se/docs/The_SPID_Algorithm_-Statistical_Protocol_IDentification.pdf (accessed 2009-04-02).

[12] B.-C. Park, Y. J. Win, M.-S. Kim, and J. W. Hong., “Towards Automated Application Signature Generation for Traffic Identification,” *NOMS: Network Operations and Management Symposium*,2008.

[13] G.Szabo, D. Orincsay, S. Malomsoky, and I. Szabo, “On the Validation of Traffic Classification Algorithms,” *PAM: Passive and Active Measurement Conference*, 2008.

[14] S. Kullback and R. A. Leibler, “On information and sufficiency,” *Annals of Mathematical Statistics*, vol. 22, pp. 49–86, 1951.

[15] Y. Zhang and V. Paxson, “Detecting Backdoors,” in *USENIX Security Symposium*, 2000.

[16] W. John, S. Tafvelin, and T. Olovsson, “Passive Internet Measurement: Overview and Guidelines based on Experiences,” *Computer Communications*, vol. 33, no. 5, 2010.

[17] M. Dusi, W. John, and kc claffy, “Observing Routing Asymmetry in Internet Traffic,” *(accessed 2009-04-02)*, 2009