ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 6 , June 2015

Characterization And Classification of Internet Backbone Traffic

Madhusmita Panda

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

ABSTRACT: We contribute to an improved 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, which reveals protocol-specific features and provides a systematic survey of packet header anomalies. The packet-level analysis is followed by a characterization on the flow-level, where packets are correlated according to their communication endpoints. We propose a method and accompanying metrics to assess routing symmetry on a flow-level based on passive measurements. This method will help to improve traffic analysis techniques. We used the method on our data, and the results suggest that routing symmetry is uncommon on non edge Internet links. We then confirm the predominance of TCP as the transport protocol in backbone traffic. However, we observe an increase of UDP traffic during the last few years, which we attribute to P2P signalling traffic. We also analyze further flow characteristics such as connection establishment and termination behaviour, which reveals differences among traffic from various classes of applications. These results show that there is a need to make a more detailed analysis, i.e., 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. This provides a rough traffic decomposition in a privacy sensitive way. Second, we present a classification method for fine-grained protocol identification by utilizing statistical packet and flow features. Preliminary results indicate that this method is capable of accurate classification in a simple and efficient way. We conclude the thesis by discussing limitations in current Internet measurement research. Considering the role of the Internet as a critical infrastructure of global importance, a detailed understanding of Internet traffic is essential. This thesis presents methods and results contributing additional perspectives on global Internet characteristics at different levels of granularity.

KEYWORDS: Internet, 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)

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 6 , June 2015

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 [1]. 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 [2]. 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 [3]. 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.

A. INTERNET MEASUREMENT

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

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 6 , June 2015

(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:

Scientific purpose: Analysis of actual Internet traffic provides much needed input for scientific simulation and modeling. Ongoing measurements will also reveal longitudinal trends and changes in the usage of network applications and protocols, and thus foster improvement and development of network protocols and services. Finally, security measures should ideally be based on a profound understanding of traffic properties and should rely on fast and reliable methods to detect unwanted traffic and network anomalies. We therefore consider modern, real-life datasets vital for the network research and development community in order to be able to react to changes in traffic properties and behaviour (for both benign and malicious reasons) in a timely fashion.

Operational purpose: While traditional network management tools based on SNMP or Net flow mainly cover critical operational requirements for ISPs, such as troubleshooting and provisioning, more advanced traffic engineering tasks (such as QoS measures and traffic shaping) often rely on classification tools and techniques based on packet-level data [15]. Not only the development, but also the validation of these techniques requires modern traffic traces collected by measurement infrastructures. Internet measurements can also be the basis for refinement of network design and provisioning, design of robust protocols and infrastructure, and improvement of network performance and accounting. Furthermore, Internet measurements reflecting network behaviour as seen "in the wild" support security measures, such as refinement of rule sets for traffic filters, firewalls, and network intrusion detection systems.

Legal purpose: Monitoring and measurement of Internet traffic are also of increasing legal relevance, as manifested in the recently ratified data retention directive of the European Union, requiring communication providers to retain connection data for periods of up to two years with the purpose of enabling network forensics. Implementations of these types of regulations can directly benefit from the achievements of the Internet measurement community, offering experiences in the non-trivial task of efficient collection and analysis of large amounts of traffic. However, such privacy sensitive regulations also evoke discussions about their ethical implications.

B. INTERNET MEASUREMENT APPROACHES

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.

Software-based vs. hardware-based measurement: Passive measurement tools based on software modify operating systems and device drivers on network hosts to obtain copies of network packets (e.g., BSD packet filter). In contrast, hardware based methods are designed specifically for collecting and processing network traffic on high-speed links such as an Internet backbone. Custom-built hardware collects traffic directly on the physical links5 (e.g., by using optical splitters) or on network interfaces (e.g., mirrored router ports). Specifically, for our measurements we used a hardware-based measurement infrastructure applying optical splitters and Endace DAG cards capable of collecting unsampled, complete packet traces on links with
transmission speeds of up to 10 Gbit/s.

Online vs. offline processing: Online processing refers to immediate processing of network data in "real time", which is essential for applications such as traffic filters and intrusion detection systems. Offline processing, on the other hand, is performed on network data after it is stored on a data medium. Offline processing is not time critical and offers the possibility to process, compare, and validate network traffic collected at different times or different locations. Furthermore, stored network data can be reanalysed on the basis of different criteria. Because of these advantages, we chose offline processing for the packet and flow-level characterization presented in this thesis, since they include complex and time-consuming analysis.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 6 , June 2015

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 (e.g., flow size distribution [4]) to detailed features of single packets and flows [5]. 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 [6], some of our results revise or update previous studies that are based on outdated data sets collected years ago. Most results in this thesis are based on contemporary data from a previously unstudied measurement location on the Internet and hence contribute to a global picture of current Internet traffic characteristics.

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 [7]. Our packet analysis furthermore includes a systematic listing of packet header anomalies together with their frequencies as seen "in the wild" 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 [8–10]. 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 behaving "in the wild". These results enable the Internet community to see 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.

A. TRAFFIC CLASSIFICATION

We define traffic classification as the analysis of Internet data resulting in a decomposition of the traffic according to network applications/application layer protocols or classes thereof (e.g., bulk, interactive, WWW, etc. [11]). In other words, the goal of traffic classification is to understand the type of traffic carried on Internet links [12–14]. Traffic classification results can be useful for traffic management purposes (such as QoS and traffic shaping mechanisms [15]) and traffic engineering purposes (such as optimization of network design and resource provisioning). Furthermore, understanding the type of traffic carried on networks supports security monitoring by facilitating the detection of illicit traffic, such as network attacks and other security violations. Modern firewalls, NAT boxes, and Intrusion Detection Systems (IDSs) need to be able to reliably classify network protocols in order to implement fine grained and secure access policies. Apart from the apparent interest of operators and researchers in understanding trends and changes in network usage, there have also been a number of political and legal discussions about Internet usage, further highlighting the need for accurate traffic classification methods. These political discussions include the ongoing debate between intellectual property representatives1 and the P2P file sharing community [16, 17]. There are also network neutrality discussions between Internet Service Providers (ISPs) and content providers [15]. Historically, network applications have been designed to use well-known port numbers to communicate with servers or peers, making traffic classification relatively straightforward. However, in the early 2000s, 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 [11]. Since then, there has been an ongoing arms race between application developers trying to

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 6 , June 2015

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 .

B. INTERNET TRAFFIC APPROACH

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

Traffic characteristics

Traffic engineering presupposes good knowledge about Internet traffic behaviour as well as methods and tools for network performance measurements. Many traffic engineering methods need as input a traffic matrix describing the demand between each pair of nodes in the network.

Understanding internet traffic behaviour is essential for all aspects of network design and operation

- Component design
- Protocol design
- Provisioning
- Management
- Modeling and simulation

The packet count 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
-

Definition of self similar process: Self-similar processes are the simplest way to model processes with long-range dependence correlations that persist (do not degenerate) across large time scales

The autocorrelation function r(k) of a process (statistical measure of the relationship, if any, between a random variable and itself, at different time lags)with long-range dependence is not summable:

- Sr(k) = inf.
- r(k) @ k^{-b} as k g inf. for 0 < b < 1
  - Autocorrelation function follows a power law
  - Slower decay than exponential process
- Power spectrum is hyperbolic rising to inf. at freq. 0
- If Sr(k) < inf. then you have short-range dependence

Consider a zero-mean stationary time series $X = (X_t; t = 1,2,3,\ldots)$, we define the *m*-aggregated series $X^{(m)} = (X_k^{(m)}; k = 1,2,3,\ldots)$ by summing X over blocks of size m. We say X is *H-self-similar* if for all positive *m*, $X^{(m)}$ has the same distribution as X rescaled by m^H.

If X is *H*-self-similar, it has the same autocorrelation function r(k) as the series $X^{(m)}$ for all *m*. This is actually *distributional* self-similarity.

Degree of self-similarity is expressed as the speed of decay of series autocorrelation function using the Hurst parameter

- H = 1 - b /2
- For SS series with LRD, ½ < H < 1
- Degree of SS and LRD increases as H to1

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 6 , June 2015

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

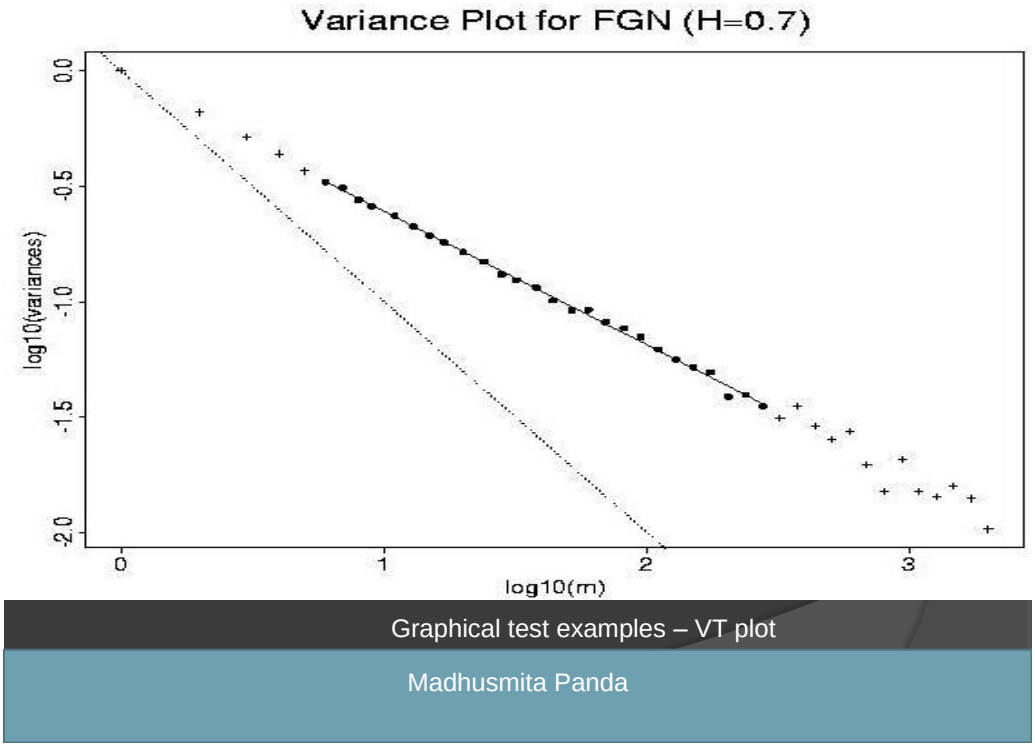

Figure 2.1 Graphical test examples – VT plot

III. DESCRIPTION OF THE NETWORKS MEASURED

GigaSUNET:

The first measurement traces we analysed were collected on the previous generation of the SUNET backbone network, called GigaSUNET. GigaSUNET was officially in operation until January 2007, when it was replaced by the current generation, called Opto-SUNET

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 6 , June 2015

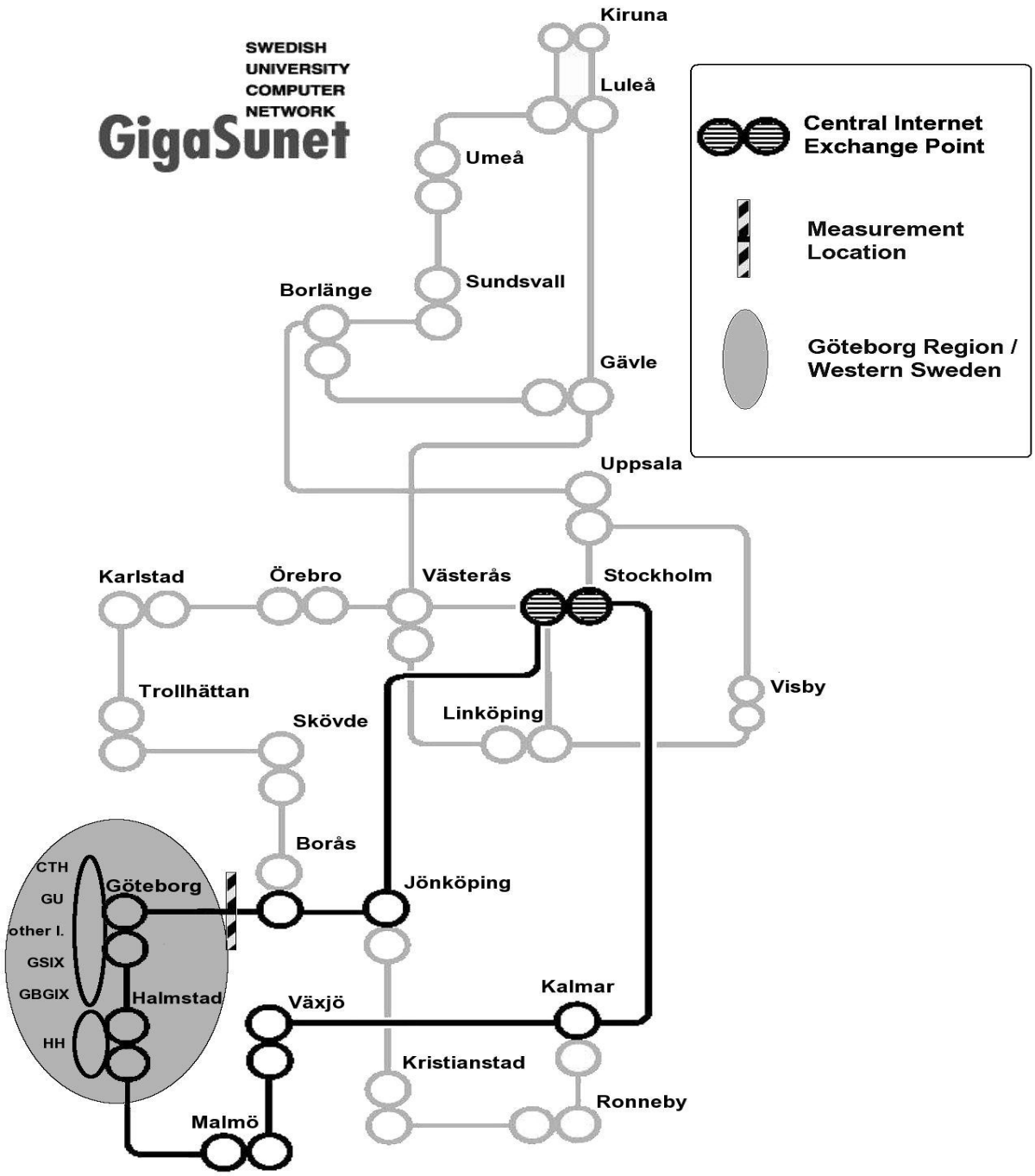

Figure 3.1: Internal GigaSUNET topology

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 6 , June 2015

IV. EXPERIMENTAL RESULTS

A. 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 in Spring 2006 to provide a summary of current protocol usage including comparisons to prior studies. The analysis confirmed that IP options and Explicit Congestion Notification (ECN) are virtually not applied. On GigaSUNET, we observed minor fractions of fragmented IP traffic (0.06%), with UDP accounting for a majority of the fragments. The latter observation stems from increased deployment of TCP Path MTU Discovery, which we showed to be dominating. Regarding packet size distribution, three findings should be noted: (i) we found packet size distribution on GigaSUNET to be bimodal, i.e., most packets were either small (44% between 40 and 100 byte) or close to the Ethernet Maximum Transmission Unit (MTU) size (37% between 1400 and 1500 byte). Earlier measurements (up to 2002) on backbone links , and also more recent wide-area measurements in China during 2006 reported of substantial fractions of packets with default datagram sizes (i.e., 576 bytes [134]); (ii) in our data from 2006, IP packet lengths of 628 bytes were even more common (1.8%) than the default datagram size (<1%). We identified these packets to be artifacts of a then popular P2P application (Gnutella [135]); (iii) we do not see any jumbo packets except for BGP updates between routers and one single custom application optimized for bulk transfer. We furthermore identified additional headers introduced by VPN as one cause for the otherwise rare occurrence of IP fragmentation, which should advise application developers to use smaller MSS values.

Collection of Traces

The traffic traces have been collected on the outermost part of an SDH ring running Packet over SONET (PoS). The traffic passing the ring to (outgoing) and from (incoming) the Internet is primarily routed via our tapped links. This expected behaviour is confirmed by SNMP statistics showing a difference of almost an order of magnitude between the tapped link and the protection link. Simplified, we regard the measurements to be taken on links between the region of Goteborg, including exchange traffic with the regional access point, and the rest of the Internet.

On the two OC-192 links (two directions) we use optical splitters attached to two Endace DAG6.2SE cards. The DAG cards captured the first 120 bytes of each frame to ensure that the entire network and transport header information is preserved. The data collection was performed between the 7th of April 2006, 2AM and the 26th of April 2006, 10AM. During this period, we simultaneously for both directions collected four traces of 20 minutes each day at identical times. The times (2AM, 10AM, 2PM, 8PM) were chosen to cover business, nonbusiness and nighttime hours. Due to measurement errors in one direction at four occasions we have excluded these traces and the corresponding traces in the opposite direction.

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 4.1(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 4.1(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 4.1 (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.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 6 , June 2015

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
ICMP	0.2	0.02	0.19	0.02	0.20	0.02	0.12	0.01
ESP	0.01	0.00	0.47	0.19	0.35	0.14	0.02	0.02
GRE	0.01	0.01	0.08	0.08	0.04	0.03	0.06	0.04

(a) IPv4 Protocol Breakdown (values in %)

OUTGOING UDP

Time	Packets	Data
2PM	6.8	1.7
8PM	40.6	5.1
2AM	51.9	6.1
10AM	58.1	7.1
2PM	5.7	1.8

(a) UDP Burst (values in %)

Table 4.1: Transport Protocols

Analysis of IP Properties
IP type of service

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

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 6 , June 2015

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.

Analysis of TCP Properties

TCP Options

In an early study, Allman reported about portions of hosts applying the Window Scale (WS) and Timestamp (TS) options, both increasing from about 15% to 20% during a 15 month period from 1998 to 2000. The SACK permitted option was shown to increase even further from 7% to 40%. No numbers for hosts applying the MMS option were given. The more recent approach to quantify TCP option deployment by Pentikousis et al. in 2004 was unfortunately carried out on traces with incomplete header information. Since TCP option data was not available in these traces, their deployment had consequently to be analysed indirectly. Our results, based on traces including complete header information, show that this indirect approach yielded quite accurate results. Table 2(a) shows the deployment of the most important TCP options as fractions of the SYN and SYN/ACK segments, divided into summaries of the four times each day. The results show that MSS and SACK permitted options are widely used during connection establishment (on average 99.2% and 89.9% resp.). The positive trend of the SACK option deployment, as indicated by Allman, was obviously continued and the inferred values of Pentikousis et al. are finally confirmed. The frequent usage of the MSS option again indicates the dominance of Path MTU Discovery in TCP connections, since an advertised MSS is the precondition for this technique. The WS and TS options on the other hand are still applied to the same extent as in 2000 (17.9% and 14.5% resp.). In Table 2(b) the occurrence of TCP options with respect to all TCP segments is summarized. Around 87% of the TCP segments do not carry any options at all. Only an average of 2.9% of all segments actually applies the SACK opportunity, which was permitted by around 90% of all connections. It is interesting, that although 15.5% of the connection establishments advertise usage of the TS option, it just reappears in 9.3% of all segments. This might be caused by TCP servers not responding with the TS option set in their initial SYN/ACK. All other option kinds were observed with very low frequency.

Kind	2AM	10AM	2PM	8PM
2(MSS)	99.0%	98.7%	99.7%	99.1%
3(WS)	21.4%	18.4%	16.6%	16.5%
4(SACK perm.)	91.0%	86.6%	88.9%	89.8%
8(TS)	18.2%	15.3%	13.3%	12.8%

(a) TCP Options in SYN segments

Kind	2AM	10AM	2PM	8PM
2(MSS)	86.5%	85.2%	87.3%	88.6%
5(SACK)	3.1%	2.8%	2.9%	3.1%
8(TS)	9.7%	11.2%	9.0%	7.6%

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 6 , June 2015

19(MD5)	0.02%	0.02%	0.01%	0.01%

(b) TCP Options in all segments

Table 4. 2: TCP Option Deployment

B. CLASSIFICATION OF INTERNET TRAFFIC

Traffic classification can be applied for various purposes (e.g., traffic management, traffic engineering, security monitoring, accounting, policy enforcement) that require different classification granularity. Routing asymmetry has mainly considered an end-to-end perspective, inferred by active measurements of delay or path differences between endpoints to our knowledge; using passive measurement to quantify routing asymmetry observed on a specific link has only received tangential reference. We propose a technique that uses passive measurements to quantify the amount of traffic routed (a) symmetrically on specific network links, in terms of flows, packets and bytes. Using passively captured network data, the Flow-Based Symmetry Estimator (FSE) method provides an effective way to exclude traffic that is canonically asymmetric, such as ICMP traffic or non productive TCP background radiation, allowing a fair comparison of routing symmetry across different links with substantially different traffic decomposition. Knowledge of the fraction of symmetric flows on specific links is especially important to traffic analysis and characterization tasks, which are often performed on data collected on single measurement points. Researchers and developers often embed an assumption of traffic symmetry in tools and analyses, an assumption only safe for stub access links, otherwise quite harmful. We wanted to provide the community with a technique and accompanying open source tool for measuring flow symmetry, as well as raise awareness about macroscopic symmetry characteristics by providing statistics from running such tools over a variety of data. We evaluated our technique on traffic traces from four varied locations (Tier-2 to Tier-1 backbone) in two countries (USA and Sweden) over a period of four years (from 2006 till 2009), to provide a baseline global data set on routing symmetry. Such data sets will allow tracking of macroscopic Internet trends. Our main contributions are: (i) a simple method to assess and fairly compare routing symmetry on specific links (ii) an open source tool for analyzing flow symmetry based on our method and (iii) symmetry statistics for a large heterogeneous set of network traces method and (iii) symmetry statistics for a large heterogeneous set of network traces.

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.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 6 , June 2015

P2P traffic

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 5.2 and 5.3 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.

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 5.1: 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 5.2: P2P Range (Link Location)

Geo Location	Year	Range of P2P Volume
Europe	2005	60-80%

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 6 , June 2015

	2005	79-93%
North America	2003	8%,10.7%
	2004	14%, 9.9%
	2006	21-35%
Asia	2002	21.5%
	2005	1.34% (port-based)
	2008	1.29% (port-based)

Table5. 3: P2P Range (Geographic Location)

Classification of Internet Traffic based on Statistical Features

Proposed Heuristics

For our data the thresholds used were derived empirically through experiments on a number of traces. In the following list of heuristics, *(K)* (Karagiannis) or *(P)* (Perenyi) indicate by which previous method the heuristic was inspired, while *(J)* (John) marks newly introduced rules.

Verification of the proposed Heuristics

To verify the proposed adjustments, we classified our backbone data by each of the three sets of heuristics. For each flow, a bitmask was set in a database according to matching rules. This method allowed us to analyze intersections between the three approaches separately - meaning flows marked as P2P traffic by either one, two or all three of the approaches. The results are illustrated by the Venn diagrams in fig.1, presenting connection counts (a) and amount of data (b) in absolute numbers. The three circles represent P2P flows classified by the different rule-sets (Karagiannis left, Perenyi right, new proposal beneath). The following paragraphs will discuss the different intersections (IS I-VII), thereby motivating the proposed modifications and additions to the original approaches.

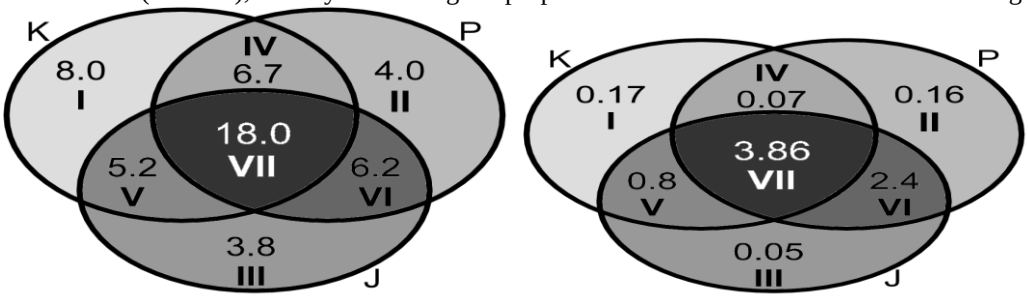

Figure5. 1: P2P traffic by Karagiannis (K), Perenyi (P) and new proposal (J)

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 6 , June 2015

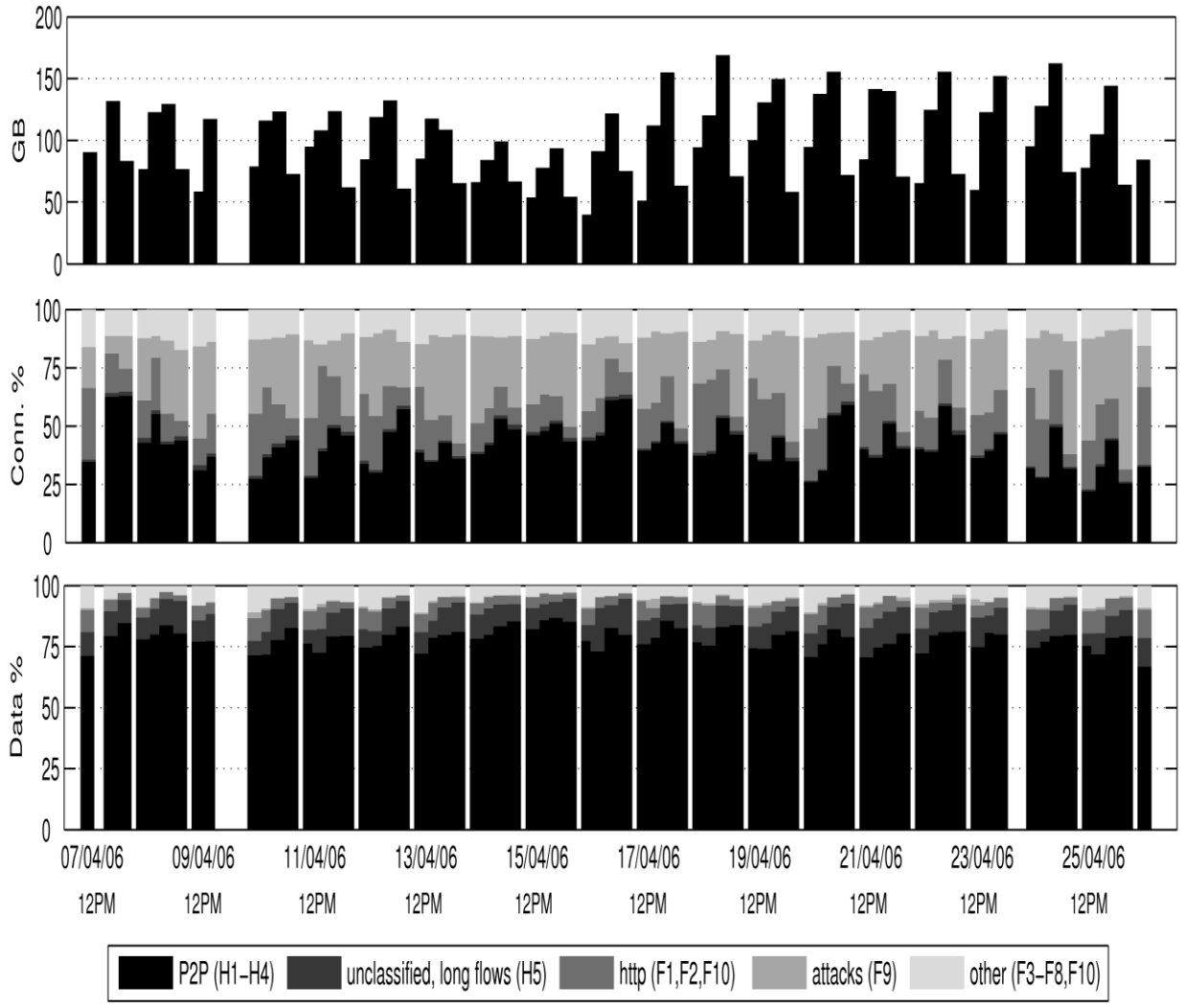

Figure5. 2: TCP data vs trace times (1st row); Appl. breakdown by #conn. (2nd); Appl. Breakdown by data carried (3rd)

Results and discussion

We finally 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 (2AM, 10AM, 2PM and 8PM). 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 [10AM, 2PM, 8PM, 2AM]. The first graph shows total amount of TCP data in GByte versus trace times. The second and third row illustrate application breakdown for the particular trace in terms of connection numbers and data volumes. 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

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 6 , June 2015

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

Considering the role of the Internet as a critical infrastructure of global importance, we claim that it is crucial for the Internet community to understand the nature and detailed behaviour of modern network traffic. A deeper understanding supports 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 are confident that we advanced the understanding of the modern Internet by presenting current characteristics of Internet traffic based on a large amount of empirical data. Furthermore, we discussed methodological aspects of passive Internet measurement and data collection.

To follow the resulting learning curve, we had to zoom out to higher levels of data aggregation, first to the flow-level and then to classes of Internet traffic. We have presented detailed Internet traffic characteristics on both packet and flow granularities. 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. Measuring actual Internet traffic revealed a great deal of behaviours that do not follow standards, which was at first somewhat unexpected for us as naive researchers expecting textbook behaviour. Almost every possible inconsistency in protocol headers and connection signalling appears "in the wild", highlighting the need for careful design and robust implementation of network applications and infrastructure to keep them resilient against the multitude of network attack types in the global Internet environment. The changing nature of Internet traffic, with new protocols and applications appearing continuously, requires ongoing revalidation of common assumptions. 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 (e.g., 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. Traffic granularities could be (i) single protocols (e.g., for security monitoring); (ii) network applications (e.g., for policy enforcement); and (iii) protocols merged into application classes (e.g., for traffic engineering and accounting), which could be inspired by categories used in existing work We have presented our experiences from passive backbone data collection by breaking the main obstacles down to separate challenges: economic, legal, ethical, operational, and technical considerations. 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

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 6 , June 2015

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] H. Kim, kc claffy, M. Fomenkov, D. Barman, M. Faloutsos, and K. Lee, "Internet Traffic Classification Demystified: Myths, Caveats, and the Best Practices," *ACM CoNEXT*, 2008.
[16] E. Bursztein, "Probabilistic Identification for Hard to Classify Protocol," in *WISTP: Workshop in Information Security Theory and Practices*, 2008.
[17] Y. Zhang and V. Paxson, "Detecting Backdoors," in *USENIX Security Symposium*, 2000.
[18] W. John, S. Tafvelin, and T. Olovsson, "Passive Internet Measurement: Overview and Guidelines based on Experiences," *Computer Communications*, vol. 33, no. 5, 2010.
[19] M. Dusi, W. John, and kc claffy, "Observing Routing Asymmetry in Internet Traffic," *(accessed 2009-04-02)*, 2009