Monitoring three National Research Networks for Eight Weeks: Observations and Implications

Demetris Antoniades, Michalis Polychronakis, Nick Nikiforakis, Evangelos P. Markatos
Institute of Computer Science Foundation for Research & Technology – Hellas
{danton,mikepo,nikifor,markatos}@ics.forth.gr
Yiannis Mitsos
GRNET
{ymitsos}@admin.grnet.gr

***Abstract*—With the advent of dynamic and elusive distributed applications such as peer-to-peer file sharing systems, network administrators find it increasingly difficult to understand the types of applications running in their networks and the amount of traffic each application produces.**

In this paper, we present measurement results from the deployment of an accurate traffic characterization application in three National Research and Education Networks for a period of two months. Our observations go beyond traffic distribution; we explore the application usage in terms of active IP addresses, the existence of IP addresses generating massive amounts of traffic, the asymmetry of incoming and outgoing traffic, and the existence of SPAM-sending mail servers.

I. Introduction

Network traffic volume increases at a rate of about 50% every year [2]. Besides the growth in traffic rates, the number of users, hosts, domains and enterprise networks connected to the Internet has been also growing explosively. These continuously increasing numbers of Internet usage and traffic volume derive from the deployment of new and massively used applications that fulfill the requirements of connected users. These applications, mostly peer-to-peer file sharing applications, are used mainly for the distribution of large files, and contribute a great volume of the Internet traffic. Understanding the usage of Internet traffic both from the perspective of different applications and the contribution of different IP addresses in the traffic mix is important for both network design and administrative tasks.

In an effort to bypass firewall and traffic shaping restrictions, new applications and protocols avoid the use of static, predefined port numbers [1], as used by conventional applications, such as HTTP and FTP, since the beginning of wide-spread networking. Instead, they use arbitrary, dynamically allocated ports, which render classification mechanisms based on port numbers highly inaccurate [11], [14]. Aiming in deeper traffic analysis, by revealing traffic patterns for different applications, we used a traffic classification application based on deep packet inspection (similar to [10], [16]), and flow-state analysis, in order to classify the monitored network traffic.

In this paper we present a multi-dimensional characterization of the traffic of three National Research and Education Networks for a period of two months. Using our traffic classification application we attribute traffic to the application that generated it and present the observations extracted from the results. We observe that peer-to-peer file sharing applications completely dominate the traffic, while HTTP flows are a significant part of the total incoming traffic. Exploring the traffic behavior over time for different applications, we observe that while HTTP traffic shows a clear diurnal cycle, the rest of the significant traffic-generating protocols do not seem to have such an observable cycle. Looking at the traffic transferred by each distinct IP address, we identify a small percentage of hosts to be responsible for more than 99% of the total traffic. The same pattern is preserved when distinct network flows are categorized by the application of origin. Finally, we take a closer look at the observed SMTP traffic, trying to identify suspicious email usage. Our results show that only around 50% of all SMTP traffic comes from valid, registered mail servers.

The paper continues as follows. Section II presents related work regarding both general and per-application traffic patterns. Section III describes the measurement environment and the data collected. An analysis of the contribution of different IP addresses in the traffic mix is presented in Section IV. Section V explores the symmetry of transferred traffic for different IP addresses and applications. Section VI studies the periodicity of different applications. In Section VII we present our findings regarding SPAM relays based on traffic pattern observations. Finally, Section VIII concludes our work.

II. Related Work

Traffic analysis has always been of great interest to the networking research community. In 1974, Kleinrock and Naylor [13] measured the network behavior of ARPANET and noted that a small number of sources were responsible for a large portion of the traffic. Claffy et al. [5] analyzed the traffic of the T1 NSFNet backbone, and presented, among other observations, that a few networks were responsible for a large percentage of the traffic. In the late nineties, researches studied network traffic patterns at the AS level [8] and presented similar observations about the network traffic contributors. Recent studies [7], [15] defined “elephants” in network traffic as the flows with high volume of traffic and long persistence in time. Though in our work we also look at heavy network traffic producers, we differ from previous work by changing our observation point and looking at the traffic from the IP(host)

level, while we also present an analysis about the different applications protocols used.

Karagianis et al. [12] showed that the usage of P2P applications continues to grow, despite several media reporting decrease due to the legal campaign in favor of copyrighted content. In the analysis provided by [6], the authors show HTTP to be the dominant client-server application protocol. The analysis also shows that about 10% of the IP addresses are responsible for 90% of the traffic. Gerber et al. [9] showed that the majority of traffic is generated by a small percentage of the total IP addresses, while they find P2P applications to dominate both incoming and outgoing traffic. On the other hand, they show diurnal behavior for both Web and P2P traffic. Our results contradict this observation, but agree with [4]. We observe diurnal behavior for HTTP traffic, while other protocols, including P2P, do not expose such behavior. The most popular P2P protocol varies according to the region and time of the study.

III. Collection Methodology

We installed a passive monitoring sensor that monitors the traffic of three National Research and Education Networks (NRENs). The sensor is located at the (common) edge of the three networks and monitors all incoming and outgoing traffic from and to the Internet. The sensor runs *appmon*, an accurate traffic classification application. *Appmon* passively monitors the traffic passing trough the monitored link and categorizes the active network flows (identified by source and destination IP addresses, source and destination port numbers and transport layer protocol) according to the application that generated them, using algorithms based on deep packet inspection. For each internal IP address, *appmon* stores both the incoming and outgoing traffic rate observed for each protocol in the last minute.We refer the reader to [3] for more information about the tool and its classification procedure.

The data presented in this paper were collected for a period of more than two months, from 6 August 2007 to 11 October 2007. During this period, *appmon* processed about 77.5 TB of data, originating from 152,320 IP addresses. Table I summarizes the volume of the traffic observed from each network during the measurement period. We observe a large difference in the number of IP addresses that received traffic in comparison to the number of IP addresses that transmitted traffic. We speculate this to be due to limited usage of the applications, inconsistent application caches (that is, a host that previously participated in the application network has now a different IP address and other hosts prompt for it), or attack traffic, such as port scanning activities or backscatter traffic. The largest percentage of the traffic, about 70%, was produced by three protocols: BitTorrent, eDonkey, and HTTP. We also noticed that file sharing P2P applications upload much more traffic than they download, while traditional client-server applications exhibit the exact opposite behavior.

IV. Bytes transferred Per IP

In this section, we explore the discrepancy among different IP addresses with respect to the network traffic transmitted or received by the corresponding hosts. Our observations show that a *small* percentage of hosts is responsible for the vast majority of the traffic.

Figure 1 plots the percentage of hosts responsible for the cumulative percentage of the incoming traffic for each network (dots). As we can see, 99% of the traffic is destined to 0.31% of the IP addresses for NREN1, 7% for NREN2 and 9% for NREN3. The observations suggest a highly skewed distribution of inbound traffic: a small percentage of IP addresses is responsible for most of the downloaded traffic. Figure 1 also shows that the same observations hold for the outgoing traffic (cross points). For all three networks, the 99% of the outgoing traffic is generated by less than 10% of the IP addresses.

Taking our analysis one step further, we look at the relationship between hosts and traffic percentage in a per-application basis. Our results show that for all protocols, less than 10% of the IP addresses are responsible for 90% of the traffic. For the most used protocols (HTTP, FTP, BitTorrent and eDonkey), the percentage of IP addresses drops to less than 5% for the 90% of both incoming and outgoing traffic.

V. Incoming and Outgoing Traffic Symmetry

In this section we try to understand whether hosts behave mainly as information producers, as information consumers, or as both. Figure 2 shows the relationship between incoming and outgoing traffic for the whole two-month period. Each point corresponds to an IP address and represents the volume of incoming traffic (x coordinate on the x-axis) and outgoing traffic (y coordinate on the y-axis) in KBytes. To exclude pathological cases, we have removed those IP addresses that had zero incoming or outgoing traffic—in most cases such behavior was the result of scanning and/or backscatter activity. Hosts that download and upload similar amounts of traffic tend to cluster around the $x = y$ diagonal line. Points well below the diagonal denote heavy consumers, while points well above the diagonal denote heavy producers.

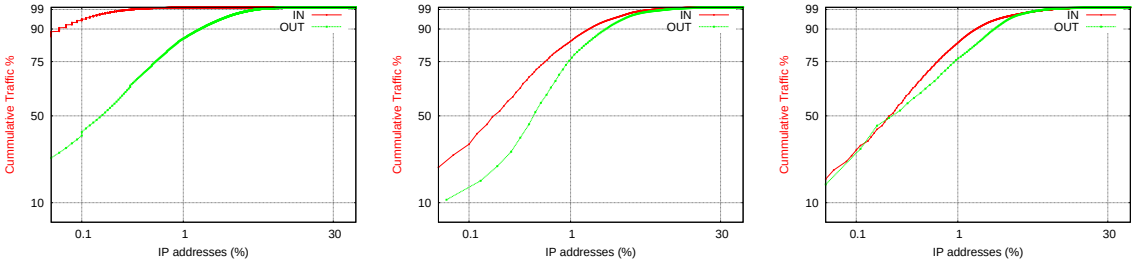
According to the results of Figure 2, about 40% of the hosts can be characterized as producers, since they transmit more than twice the traffic they receive, while 43% of the hosts behave as consumers, accepting twice the traffic they output. The remaining 17% behave both as producers and as consumers, transmitting as much (within a factor of two) traffic as they receive. Although the total traffic mix has similar percentages of consumers and producers, individual protocols are dominated by either consumers or producers. For example, HTTP has more than 53% of the hosts functioning as consumers, versus 33% producers, which, given the nature of the HTTP protocol, is expected. It is interesting to note that BitTorrent and eDonkey exhibit a behavior closer to producers. For these applications, 48% and 39% of the IP addresses act as producers, respectively, while only 25% and 10% act mainly as consumers.

VI. Traffic Periodicity

We now turn our attention to the periodicity observed in network traffic. Figure 3(a) shows the total traffic for all three NRENs for a period of one month (20 September 2007

TABLE I
TOTAL VOLUME OF TRAFFIC OBSERVED DURING THE MEASUREMENT PERIOD AND NUMBER OF MONITORED AND ACTIVE IP ADDRESSES

	TBytes Received	TBytes Transmitted	Total IP Range	Incoming Active IPs	Outgoing Active IPs
NREN1	10.5	32.9	139776	139691	11880
NREN2	1.6	1.2	4096	4096	1757
NREN3	5.1	12.0	8448	7991	1892
Total	21.9	55.5	152320	151778	15529



(a) NREN1 (b) NREN2 (c) NREN3

Fig. 1. **Cumulative percentage of incoming (red line) and outgoing (green line) traffic as a function of the percentage of the IP addresses receiving this traffic.** We see that a small portion of IP addresses is responsible for the majority of the downloaded traffic.

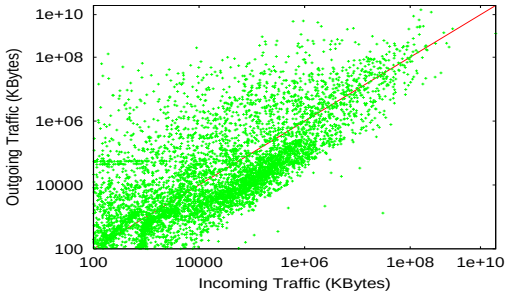


Fig. 2. **Traffic producers vs. traffic consumers.**

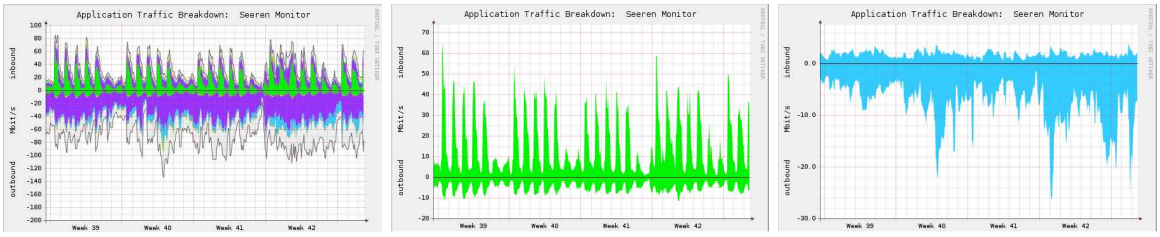
to 20 October 2007). We can clearly see that the network traffic follows a diurnal cycle. We observe five noticeable spikes per week which probably correspond to daily human activities. If we exclude the HTTP traffic (Figure 3(b)) and focus our attention on other application protocols, we see that the diurnal pattern suddenly disappears. Figures 3(c)–3(f) present the BitTorrent, eDonkey, SSH, and FTP traffic separately for the same time period, which do not exhibit any obvious diurnal cycle.

VII. SERVING OR SPAMMING?

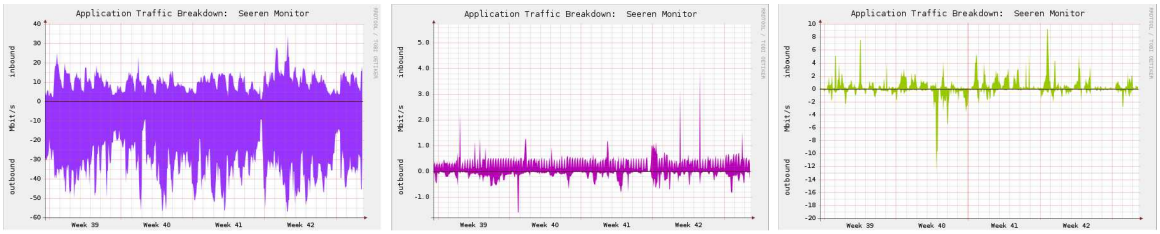
In our traces, we identified a significant number of IP addresses (4244) transferring traffic over port 25. Port 25 is used by mail servers for exchanging emails and by email clients for communicating with the mail server. Since our monitor was located at the edge of the network, the point that connects the organizations with the rest of the Internet, we did not expect to monitor sendmail traffic originating from clients towards servers, because, in most cases, both hosts were expected to be located within the organization. Even though external webmail services, such as gmail, hotmail and yahoomail, also support SMTP for use with email client software, we did not expect significant use of this functionality. Furthermore, in a proper setup, the mail server is not a user-operated machine, thus it is not expected to transfer other application traffic besides SMTP and other email-related protocols. We consider IP addresses that both receive and transmit SMTP traffic and at the same time transfer other kind of traffic, such as, BitTorrent, eDonkey or Gnutella, to be potentially infected by some kind of malware, e.g., a mail proxy utilized for sending spam.

Based on the above heuristic, we tried to count the number of hosts that are suspicious to be infected. Among the 4244 addresses, we distinguish the ones transferring both SMTP and other kind of traffic. Our results show that 3044 of these IP addresses have also exhibited activity related to at least one of the BitTorrent, eDokney, or Gnutella protocols. Taking the analysis one step further, we exclude from our list the hosts that did not produce significant SMTP traffic, i.e., we exclude IP addresses with cumulative —both incoming and outgoing— traffic of less than 1MB for the whole measurement period. The intuition behind this is that such a small amount of traffic may be due to SYN or SYN-ACK packets corresponding to scan and backscatter traffic. This step resulted to a list of 179 IP addresses that both transferred a significant number of SMTP traffic and also had traffic originating from applications not expected to be running on a server machine.

As a second step, we used reverse DNS lookup and tried to distinguish valid mail servers by the existence of MX records. Out of the 279 IP addresses, we got reverse-DNS responses for 169 IP addresses, while only 81 of them where included in the MX records. We consider these 81 IP addresses to be valid mail servers, while we mark the rest 198 as suspicious to be running rogue mail servers. These suspicious mail servers account for 54.1% of the incoming and 46.01% of the outgoing SMTP traffic.



(a) Total Traffic (b) HTTP (c) eDonkey



(d) BitTorrent (e) SSH (f) FTP

Fig. 3. **Diurnal cycles for total traffic and separate protocols.** Although total (a) and HTTP traffic (b) have a clear diurnal cycle, probably due to daily user activities, other applications however do not exhibit any obvious diurnal cycles.

VIII. Conclusions and Future Work

We have presented our initial observations from a multi-dimensional characterization of the traffic of three National Research and Education Networks for a period of two months. Using *appmon*, a per-application traffic classification application, we managed to derive traffic patterns for both the total traffic mix and for several applications separately. The analysis revealed interesting observations about today's Internet traffic. In the near future, we plan to further analyze the traffic and explore the discussed dimensions deeper, aiming to shed more light to the reasons and implications behind them.

Acknowledgments

This work was supported in part by the project CyberScope, funded by the Greek Secretariat for Research and Technology under contract number PENED 03ED440, and by the IST project LOBSTER funded by the European Union under contract number 004336. D. Antoniades, N. Nikiforakis, M. Polychronakis and Evangelos P. Markatos are also with the University of Crete.

References

[1] Internet assigned numbers authority. well known port numbers list. http://www.iana.org/assignments/port-numbers.

[2] Minnesota Internet Traffic Studies (MINTS). http://www.dtc.umn.edu/mints/home.html.

[3] D. Antoniades, M. Polychronakis, S. Antonatos, E. P. Markatos, S. Ubik, and A. Oslebo. Appmon: An application for accurate per application traffic characterization. In *Proceedings of IST Broadband Europe 2006 Conference*, December 2006.

[4] K. Claffy and N. Brownlee. Understanding internet traffic streams: Dragonflies and tortoises. *IEEE Communications Magazine*, 40(10):110–117, 2002.

[5] K. Claffy, G. Polyzos, and H. Braun. Traffic characteristics of the T1 NSFNET backbone. *INFOCOM'93. Proceedings. Twelfth Annual Joint Conference of the IEEE Computer and Communications Societies. Networking: Foundation for the Future. IEEE*, pages 885–892, 1993.

[6] T. Dang, M. Perenyi, A. Gefferth, and S. Molnar. On the Identification and Analysis of P2P Traffic Aggregation? *LECTURE NOTES IN COMPUTER SCIENCE*, 3976:606, 2006.

[7] C. Estan and G. Varghese. New directions in traffic measurement and accounting: Focusing on the elephants, ignoring the mice. *ACM Trans. Comput. Syst.*, 21(3):270–313, 2003.

[8] W. Fang and L. Peterson. Inter-AS traffic patterns and their implications. *Global Telecommunications Conference, 1999. GLOBECOM'99*, 3, 1999.

[9] A. Gerber, J. Houle, H. Nguyen, M. Roughan, and S. Sen. P2P The Gorilla in the Cable. *National Cable & Telecommunications Association (NCTA) 2003 National Show*, 2003.

[10] T. Karagiannis, A. Broido, N. Brownlee, K. Claffy, and M. Faloutsos. File-sharing in the Internet: A characterization of P2P traffic in the backbone. *University of California, Riverside, USA, Tech. Rep*, 2003.

[11] T. Karagiannis, A. Broido, N. Brownlee, K. Claffy, and M. Faloutsos. Is P2P dying or just hiding. *IEEE Globecom*, 2004.

[12] T. Karagiannis, A. Broido, and M. Faloutsos. Transport layer identification of P2P traffic. *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, pages 121–134, 2004.

[13] L. Kleinrock and W. Naylor. On measured behavior of the ARPA network. *AFIPS Conference Proceedings, National Computer Conference*, 43, 1974.

[14] A. Moore and K. Papagiannaki. Toward the Accurate Identification of Network Applications. *Passive And Active Network Measurement: 6th International Workshop, PAM 2005, Boston, MA, USA, March 31-April 1, 2005: Proceedings*, 2005.

[15] K. Papagiannaki, N. Taft, S. Bhattacharyya, P. Thiran, K. Salamatian, and C. Diot. A pragmatic definition of elephants in internet backbone traffic. *Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurment*, pages 175–176, 2002.

[16] S. Sen, O. Spatscheck, and D. Wang. Accurate, scalable in-network identification of p2p traffic using application signatures. *Proceedings of the 13th conference on World Wide Web*, pages 512–521, 2004.