Internet usage at elementary, middle and high schools: A first look at K-12 traffic from two US Georgia counties.

Robert Miller[1], Warren Matthews[2], and Constantine Dovrolis[1]

[1] Georgia Institute of Technology
robert.miller@gatech.edu,dovrolis@cc.gatech.edu
[2] JANET
warren.matthews@ja.net

Abstract. Earlier Internet traffic analysis studies have focused on enterprises [1, 6], backbone networks [2, 3], universities [5, 7], or residential traffic [4]. However, much less is known about Internet usage in the K-12 educational system (elementary, middle and high schools). In this paper, we present a first analysis of network traffic captured at two K-12 districts in the US state of Georgia, also comparing with similar traces collected at our university (Georgia Tech). An interesting point is that one of the two K-12 counties has limited Internet access capacity and it is congested during most of the workday. Further, both K-12 networks are heavily firewalled, using both port-based and content-based filters. The paper focuses on the host activity, utilization trends, user activity, application mix, flow characteristics and communication dispersion in these two K-12 networks.

1 Introduction

K-12 networks are unique for several reasons. First, they are used primarily by a very specific part of the population: children and adolescents. Second, these networks are mostly used for educational purposes, as opposed to business, research or entertainment. Third, the conventional wisdom at least is that K-12 networks are often under-provisioned in terms of Internet access capacity, experiencing congestion during most of the working day. Fourth, again based on conventional wisdom, K-12 networks are tightly controlled in terms of allowable applications and downloadable content.

Our objective in this paper is to analyze K-12 Internet traffic so that we can better understand how the Internet is used in these unique networks. Which are the dominant applications? What is the diurnal utilization pattern? Are there significant differences between say elementary schools and high schools? How does congestion affect usage, and in particular, the flow size distribution or the per-flow throughput? Further, we would like to examine the previously mentioned conventional wisdoms and understand the differences between K-12 traffic and the more often studied university traffic.

The Schools: We have collected data from two K-12 districts (counties) in the state of US Georgia: Barrow and Walton. These districts are geographically close and of similar size but they have very different Internet access capacities. Barrow is connected through a 150Mbps link to PeachNet (the education network of the state of Georgia), while Walton has a 20Mbps connection to a commercial provider. Barrow's access link has plenty of available capacity, while Walton is heavily congested during the school day.

The Barrow network is used by approximately 12,000 students and teachers at 16 schools (3 high, 4 middle and 9 elementary schools). Barrow also has 3 administrative facilities that use its network. The Walton network is used by approximately 13,000 students and teachers at 13 schools (2 high, 3 middle and 8 elementary schools). In both counties, the networks are subnetted based on schools and as such we can identify the school that each IP flow belongs to. Both networks are NAT-ted. In Barrow, our monitor is located inside the private network, and so we can identify individual hosts. In Walton, on the other hand, our monitor is located after the NAT and so we cannot identify individual hosts (even though we can still identify individual schools because each school uses a different public IP address).

To compare K-12 traffic with the more often studied university traffic, we have also collected network traces at Georgia Tech. The Georgia Tech network has several 1Gbps access links, and it is used by approximately 20,300 students and faculty. Further, Georgia Tech, as most US universities, does minimal filtering of application ports or content. While we are able to compare the K-12 traffic to the Georgia Tech data in some instances, in others we could not do so. This is mostly due to limitations on our data collection at Georgia Tech.

The Data: Data was captured using port-mirroring at the central switch of both K-12 networks. The data was stored in *nfdump* files rotated every 5 minutes.[1] In this paper, we only present nfdump data from the week of April 14-20 2008, which was a typical week for both counties in terms of usage and school operation. For Georgia Tech, we analyzed netflow data from the access router collected on September 8, 2008.

We also used a packet sniffer to extract various HTTP headers and the DNS-query field of DNS requests. That data was collected on September 8, 10 and 21, 2008 only at Walton county.

The structure of the paper is as follows. In section 2 we describe the broad characteristics of each network. In section 3 we give the breakdown of captured traffic in terms of protocols and applications. In section 4 we compare flow-level characteristics between the schools. We conclude in section 5.

2 Network Characteristics

Host Count and Activity: We first estimate the number of network-connected hosts at Barrow County. We cannot do the same for Walton because of the

[1] nfdump is a tool that collects and processes netflow data via the command line. It is part of the NfSen project: http://nfsen.sourceforge.net/

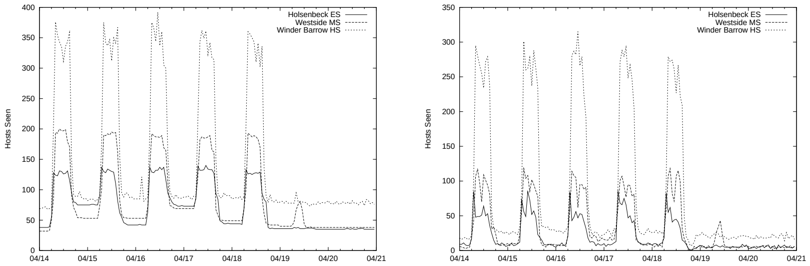
Table 1. Maximum number of hosts seen at each subnet. ES stands for elementary school, MS for middle school and HS for high school.

School	Max Hosts Seen	School	Max Hosts Seen
Barrow County	2970	ELC	38
Auburn ES	108	Russell MS	481
Bethlehem ES	143	Westside MS	205
Bramlett ES	139	Window Barrow MS	201
County Line ES	186	Apalachee HS	583
Holsenbeck ES	159	Winder Barrow HS	465
Kennedy ES	172	PLC	143
Statham ES	209	Others	32
Yargo ES	162		

previously mentioned NAT issue. Table 1 shows the maximum number of distinct hosts seen at each subnet (school) in a single day.[2] Note that the two high school networks tend to be larger in terms of hosts than middle and elementary schools.

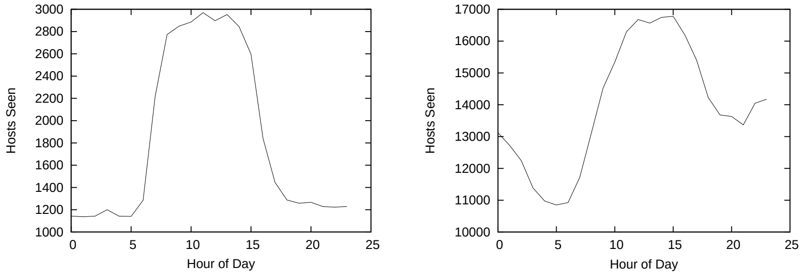
Next, we focus on the diurnal pattern of the number of hosts that are turned-on. We assume that such hosts will be generating/receiving some traffic (e.g., for network management reasons) if they are connected to the network. Figure 1(a) shows the hourly progression of hosts over the course of the week for three representative schools. We chose to display one school of each type (ES, MS, HS). As expected, the number of active hosts increases during school days, from about 7am to about 5pm. What is also interesting, however, is that a significant fraction of hosts (20% to 40%) are turned-on during evenings and weekends. Most likely, not all of these machines are servers. This indicates that these machines have been left on during off hours and weekends, probably without reason.

[2] Russell Middle School has a few special machines (mostly servers) assigned to its subnet, causing some unusual results. The "Others" category represents machines that do not belong to any school in particular.



(a) Number of active hosts (b) Hosts with HTTP/S activity

Fig. 1. Hourly activity of Barrow hosts during a week.



(a) Barrow County Apr/16/08 (b) Georgia Tech Sep/08/08

Fig. 2. Number of active hosts

We are also interested in the number of visible hosts due to user-initiated activity. It is not easy to infer whether a machine is currently used by a human or not. The heuristic that we use is to detect whether a host generates or receives any HTTP or HTTPS (denoted as HTTP/S) traffic during that time period, assuming that most (but clearly not all) HTTP traffic is due to user-initiated web browsing. Figure 1(b) shows the number of HTTP/S-active hosts during a week, in hourly intervals. Note that the activity at Holsenbeck ES and Westside MS is close to zero during the evening hours, but the same is not true for Winder Barrow HS. This may be due to hosts running Web sessions with periodic page-refreshes.

It is interesting to compare host activity between a K-12 and a university network. Figures 2(a) and 2(b) show the number of hosts seen at Barrow county and at the Georgia Tech network in hourly intervals over the course of a day. We see a very different pattern. At Georgia Tech, the number of active hosts remains at high levels during the evening hours, until midnight or so, as many students and faculty work during after-hours at the school or from home. Also, we see again that a large fraction of hosts (about 65%) remains turned-on and network-active during the evenings. This is an issue that large organizations will have to address if we are to reduce power dissipation and energy demands.

Network Utilization: Figure 3 shows how the utilization varies over the course of a weekday at Walton, Barrow and Georgia Tech, in five-minute intervals. We show two curves in each graph, for incoming and outgoing traffic. Walton is congested, with over 90% utilization of its access link, from about 8am until about 3-4pm. Further evidence of Walton's congestion is evident in the RTTs (not shown here) between Georgia Tech and the monitoring machine at Walton. During peak hours, the RTT reaches 300ms, up from around 8ms during off hours. The peak load at Barrow is about 45Mbps, much below the 150Mbps capacity. During peak hours in Barrow county we do not see significant RTT fluctuation, with the RTT staying around 2ms over the course of the day. Also, the two counties are mostly consumers of Internet traffic; the outgoing

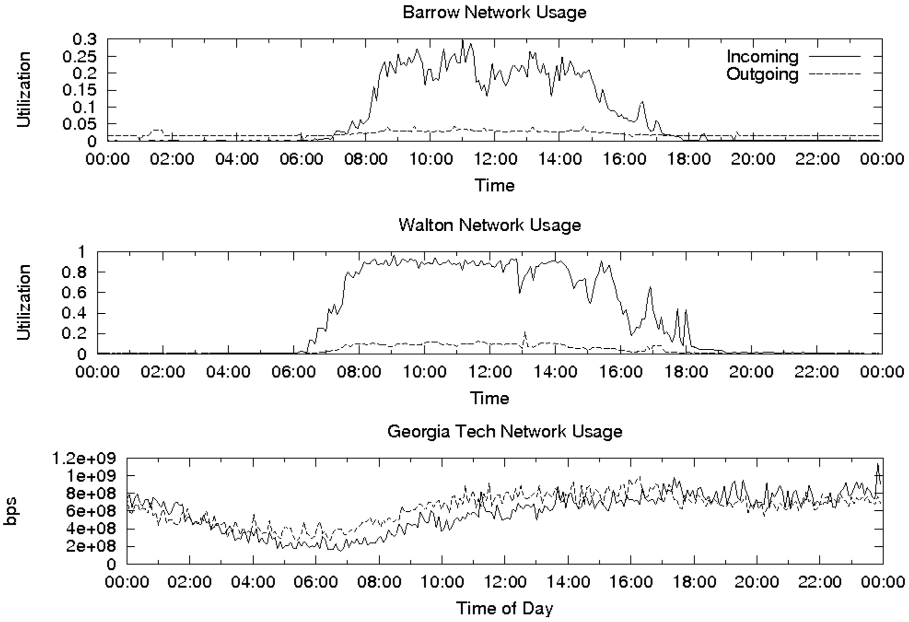


Fig. 3. Network load variations during a working day

traffic rate (mostly DNS and HTTP requests as well as outgoing email) is only 11% of the incoming rate.

Georgia Tech's diurnal usage pattern, on the other hand, is very different. It generates an almost symmetric traffic load between the incoming and outgoing directions. As we will see in the next section, this is probably because Georgia Tech acts as a significant HTTP content provider (through multiple research and software distribution servers) and because it allows the activity of peer-to-peer (p2p) applications. Further, the load variations at Georgia Tech during the day are much smaller than in K-12 networks: the minimum traffic load (200Mbps at about 6am) is 20% of the maximum traffic load (1Gbps in afternoons and evenings) [3]. On the other hand, the traffic load at K-12 networks is almost zero in the late evening hours.

3 Traffic Characteristics

Protocol and Application Breakdown: We first examine the breakdown of traffic in terms of transport protocol. The main protocols in both K-12 networks are TCP and UDP covering together more than 99% of the bytes. TCP dominates the transport layer, with almost 95-96% of the packets and 97-100% of the bytes

[3] Georgia Tech's true capacity was undisclosed. Therefore we present only the measured bps instead of utilization.

Table 2. Application layer breakdown (outgoing data/incoming data)

	HTTP	HTTPS	RTMP	SMTP	RTSP	Unknown
Barrow Packets	74.1%/79.5%	8.1%/6.9%	5.0%/5.4%	6.5%/3.2%	2.4%/2.5%	2.4%/1.7%
Barrow Bytes	66.5%/82.9%	12.2%/4.6%	1.6%/6.7%	11.4%/0.9%	0.8%/2.7%	3.4%/1.4%
Walton Packets	75.2%/79.6%	6.8%/6.1%	2.7%/1.9%	6.7%/4.2%	1.4%/1.9%	6.6%/5.6%
Walton Bytes	74.2%/85.9%	10.9%/5.3%	1.0%/4.2%	8.9%/0.3%	0.5%/2.2%	4.1%/1.9%

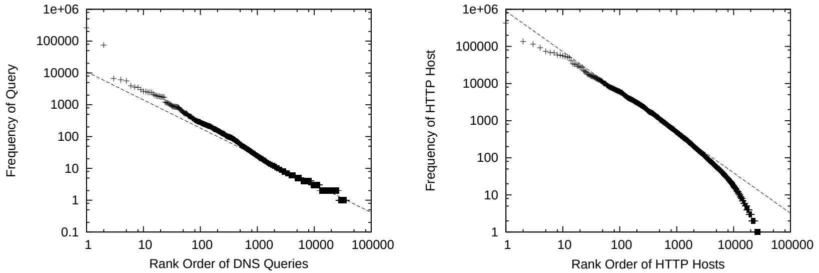
	Unknown	HTTP	rsync	DNS	NNTP	RTMP
Georgia Tech Packets	42.8%/37.5%	36.2%/37.3%	5.8%/3.6%	3.1%/4.9%	1.0%/1.8%	1.0%/2.0%
Georgia Tech Bytes	43.9%/33.2%	36.4%/49.3%	10.5%/0.6%	0.5%/0.7%	0.1%/3.9%	0.1%/3.82%

at Walton county. The UDP percentages are much higher at Barrow, but this is due to a single IP address at Russell MS that multicasts a CNN video stream using UDP. If we exclude Russell from the analysis, Barrow is also dominated by TCP traffic, with similar numbers as Walton. On the other hand, the data from Georgia Tech shows a significantly larger fraction of UDP traffic, about 11-13% of packets and 4-5% of bytes.

We next examine the application breakdown. We use a simple port-based classifier. It is well-known that this classifier is inaccurate because it fails to detect p2p or other applications that do not use well-known port numbers [5]. However, as will be shown next, this is not an issue for these K-12 networks because they block most traffic, excluding traffic from well-recognized port numbers such as HTTP or HTTPS.

Table 2 shows the application breakdown at the two K-12 networks, as well as at Georgia Tech, both for outgoing and incoming traffic. The major application-layer protocols at Barrow and Walton are HTTP/S, SMTP (email), RTMP (Real-Time Messaging Protocol, a proprietary Adobe protocol for media streaming using a Flash player) and RTSP (Real-Time Streaming Protocol, used by media clients to control remote media servers with VCR-like capabilities). HTTP/S dominates, with about 80-90% of the packets and bytes in both directions. Of course we should be aware that some applications use the HTTP/S port numbers today to “disguise” as Web browsing. Unfortunately, we have no way to detect such applications. It is interesting that the RTMP/RTSP percentages are higher for Barrow than Walton. This may be due to the heavy congestion at Walton. Streaming is more sensitive to congestion, and if streaming applications do not perform well, people would use them less frequently. Finally, note that the percentage of unidentified traffic is quite low, typically less than 5% of the bytes. This is not surprising, given that the administrators at the two K-12 networks block all ports except those that are explicitly white-listed.

The application breakdown is very different at Georgia Tech. In that case, the percentage of “Unknown” traffic is significant, 35-45% of the packets/bytes in both directions, with slightly more outgoing traffic. Although we cannot be certain using port-based classification, we expect that most of that traffic is generated by p2p applications such as BitTorrent. HTTP/S generates roughly the same traffic volume as “Unknown” even though the fraction of incoming



(a) Requested DNS domains (b) Requested HTTP HOST fields

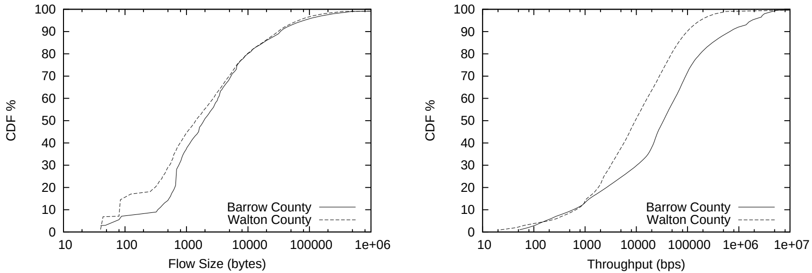
Fig. 4. Rank-order distributions on log-log scale with the associated regression curves

traffic in bytes is significantly higher than outgoing traffic. Other significant protocols at Georgia Tech are rsync (synchronization of remote file systems), DNS, NNTP and RTMP.

DNS Requests: We have also captured DNS-requests (only at Walton county) in order to characterize the domains that K-12 hosts request most frequently. Two domains were the most popular by far: walton.k12.ga.us (about 43% of the requests) and akamai.net (about 12%). This is not surprising. The Walton domain is so popular because it is probably the default web page at many hosts in that network. Akamai is the largest CDN and the web pages of their customers includes objects with Akamai DNS names. Other popular DNS domains are Google, Yahoo, AOL, MSN, Photobucket, the advertising domains llnwd.net and doubleclick.net, and nsatc.com, a domain that powers many of Microsoft's services. The top-10 domains requested capture 61% of the total requests, while the top-100 domains capture 72% of the requests. This implies that a significant fraction of the requested domains are at the tail of the distribution. Indeed, Figure 4(a) shows the frequency-versus-rank plot, in log-log scale, for the requested DNS domains. The linear trend indicates a Zipf distribution, with exponent -0.88 and R^2=96%.

HTTP Headers: We also collected HTTP headers at Walton, focusing on the Host, Content-Length and Content-type fields of the HTTP header. When we examined the median content length downloaded at Walton over the course of a day, we noticed a significant increase in the size of downloaded HTTP objects shortly after midnight. This may be due to automated software updates. In terms of Content-type, the most popular types are gif and jpeg images, followed by html, javascript and flash. We plan to compare these measurements with Barrow, when we become able to collect HTTP headers from that network.

The HTTP Host field allows us to measure the most popular Web servers the users of these K-12 networks request. The three most popular servers are Windows updates, Trend Micro, and Google. Trend Micro is an anti-virus company. The top-10 servers in the list make up 25% of all the HTTP requests, while the top-100 hosts make up 52%. We have also examined how these distributions dif-



(a) Flow size distribution (b) Per-flow throughput distribution

Fig. 5. Flow size and throughput distributions at Barrow and Walton

fer between elementary schools and high schools. In the former, we see a strong presence on websites hosting educational games. In high schools, Google and its various services dominate. We have also examined the popularity distribution of HTTP servers (see Figure 4(b)) and it also follows a Zipf distribution, but with a truncated tail. If we only consider the top 2000 HTTP servers, we get a much better fit to the Zipf model (exponent=-1.09, R^2=99%).

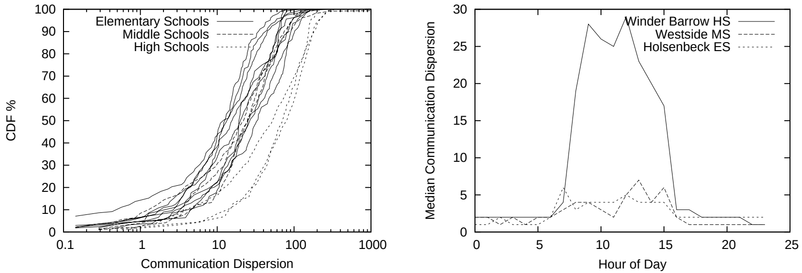
4 Flow Characteristics

In this section, we focus on the flow size distribution and the per-flow throughput in the two K-12 networks. In particular, we are interested in examining how congestion at Walton affects these two important flow characteristics.

Flow Sizes and Throughputs: Figure 5(a) shows the flow size distribution for Barrow and Walton during a working day. It is interesting that the two distributions are very similar, especially for larger flow sizes (more than 10KB) despite the fact that Walton experiences severe congestion. This observation implies that users do not react to congestion by downloading smaller files, as one may expect. Instead, it seems that they download the same files that they would download if they had more capacity.

The difference between the two distributions in smaller flow sizes, however, may be a result of congestion. In detail, we observe that the fraction of small flows (less than 10KB) is higher at Walton. This may be due to aborted flows: users often abort a transfer when it takes too long to start (due to packet losses in the TCP connection establishment or slow-start phase, for instance). The increased frequency of very small flows, compared to Barrow, is an interesting difference that we plan to further investigate.

We also examined how Walton's flow size distribution varies during the busy hours of a 9-hour working day (from 7am till 4pm). However, this distribution showed that Walton had very similar flow sizes over the course of the work day, which is not surprising given that the network is congested during this entire 9-hour period.



(a) Fan-in distributions at Barrow (b) Median fan-in at three Barrow schools on April 16, 2008

Fig. 6. Communication dispersion measured based on HTTP fan-in

Another interesting characteristic is the per-flow throughput distribution. Of course we expect much lower throughput at Walton than Barrow. Indeed, Figure 5(b) shows the corresponding distribution functions. Note that the median throughput is 33.8 kbps at Barrow and 9.5 kbps at Walton. The 90-th percentile of the throughput distribution, which may be more indicative of large-transfers, is 665.8 kbps at Barrow and only 97.7 kbps at Walton. Looking at Walton's throughput variation over the course of a day again showed us that the distribution changes very little over the course of a school day.

In summary, the results of this section indicate that congestion affects the per-flow throughput but not the size distribution of downloaded files. It is possible that users adapt to congestion by downloading files in the background, or by simply being more patient as they browse the Web. We plan to investigate this issue in more depth in the future.

Communication Dispersion: Another interesting aspect of traffic analysis is the number of outside hosts that each internal host at these K-12 networks communicates with. Here, we are primarily interested in HTTP traffic and in the incoming direction of traffic (mostly downloads). Specifically, for each host at Barrow we count the number of external HTTP servers that send traffic to that host. We refer to that number as the "fan-in" of that host. We cannot do the same for Walton due to the previously discussed NAT issue.

Figure 6(a) shows the fan-in distribution for each school within Barrow county. We label the curves based on the type of school. It is interesting that high schools have significantly higher fan-in than middle or elementary schools. This difference may be indicative of the larger diversity of content and sites that high school students (mostly adolescents) prefer, compared to the younger students (mostly children) at elementary and middle schools. This difference is further illustrated in Figure 6(b), where the median fan-in of three representative schools is shown as function of time. Note that there are no statistically significant differences between the ES and the MS, but the HS fan-in is 5-6 times larger.

5 Ongoing Work

We have recently started collecting data from more Georgia counties. We plan to expand our analysis in terms of the number of schools and the duration of the study. We are also trying to further understand the effects of congestion on user behavior and application performance.

Acknowledgments

The authors would like to thank Ed Morrison, John St.Clair and Kevin McCage at Barrow County Schools and Todd Antwine and Jon Graves at Walton County Schools for hosting the monitoring servers. We'd also like to thank Claudia Huff and the F3 group for their ongoing support and assistance. The Georgia Tech data used en this work were made available by the Georgia Tech Research Network Operations Center (www.rnoc.gatech.edu).

References

1. William Aiello, Charles Kalmanek, Patrick McDaniel, Subhabrata Sen, Oliver Spatscheck, and Jacobus Van der Merwe. Analysis of communities of interest in data networks. In *Passive and Active Network Measurement*, 2005.
2. Marina Fomenkov, Ken Keys, David Moore, and K. Claffy. Longitudinal study of Internet traffic in 1998-2003. In *ACM International Conference Proceeding Series*, 2004.
3. Chuck Fraleigh, Sue Moon, and Bryan Lyles. Packet-level traffic measurements from the Sprint IP backbone. In *IEEE Network*, 2003.
4. Kensuke Fukuda, Kenjiro Cho, and Hiroshi Esaki. The impact of residential broadband traffic on Japanese ISP backbones. In *ACM SIGCOMM*, 2005.
5. Thomas Karagiannis, Andrew Broido, Michalis Faloutsos, and Kc claffy. Transport layer identification of P2P traffic. In *ACM SIGCOMM*, 2004.
6. Ruoming Pang, Mark Allman, Mike Bennett, Jason Lee, Vern Paxson, and Brian Tierney. A first look at modern enterprise traffic. In *Internet Measurement Conference*, 2005.
7. F. Donelson Smith, Felix Hernandez Campos, Kevin Jeffay, and David Ott. What TCP/IP protocol headers can tell us about the web. In *ACM SIGMETRICS*, 2001.