Journal of Computer Science 10 (1): 99-105, 2014
ISSN: 1549-3636

doi:10.3844/jcssp.2014.99.105 Published Online 10 (1) 2014 (http://www.thescipub.com/jcs.toc)

ANALYSIS OF INTERNET TRAFFIC IN EDUCATIONAL NETWORK BASED ON USERS' PREFERENCES

Mustafa M.H. Ibrahim, Mohd Hasbullah Omar, Adib M. Monzer Habbal and Khuzairi Mohd Zaini

InterNetWorks Research Lab, School of Computing,
Universiti Utara Malaysia, 06010 UUM Sintok, Kedah, Malaysia

Received 2013-03-04; Revised 2013-07-10; Accepted 2013-11-12

ABSTRACT

The demand for Internet services and network resources in Educational networks are increasing rapidly. Specifically, the revolution of web 2.0 "also referred to as the Read-Write Web" has changed the way of information exchange and distribution. Although web 2.0 has gained attraction in all sectors of the education industry, but it results in high-traffic loads on networks which often leads to the Internet users' dissatisfaction. Therefore, analyzing Internet traffic becomes an urgent need to provide high-quality service, monitoring bandwidth usage. In this study, we focus on analyzing the Internet traffic in Universiti Utara Malaysia (UUM) main campus. We performed measurement analysis form the application level characteristics based on users' preferences. A total of three methodological steps are carried out to meet the objective of this study namely data collection, data analysis and data presentation. The finding shows that social networks are the most web applications visited in UUM. These findings lead to facilitate the enhancement of Educational network performance and Internet bandwidth strategies.

Keywords: Web 2.0, Traffic Characteristics, Educational Network, Traffic Measurement

1. INTRODUCTION

Web 2.0 is a concept that takes the network as a platform for information sharing and collaboration on the World Wide Web. It goes beyond the one-way provision of downloadable content by allowing users to become contributors. Web 2.0 has inspired intense and growing interest, particularly as wikis, weblogs (blogs), Really Simple Syndication (RSS) feeds, social networking sites and peer-to-peer media-sharing applications. The attractive features of web 2.0 tools offer higher education institutions opportunities to move away from the last century's highly centralized, industrial model of learning and toward individual learner empowerment through designs that focus on collaborative, networked interaction (Rogers *et al*., 2007; Sims, 2006; Sheely, 2006).

Universities and higher institutions keep updating their Internet services. However, web 2.0 applications are growing faster and faster. In addition, the extensive usage of these applications consumes high bandwidth, leading to network congestion and performance degradation. Therefore, it becomes imperative to study the utilization of network resources and the distribution of Internet traffic flows in educational networks. This is the task of network administrators who are searching for the best performance that their network can perform. There are three major components should be measured to detect the performance of applications; First one is to monitor and analyze network bandwidth also called throughput; Second one is latency also called delay and the third one is Internet traffic characterization, which is the point of focus for this study (Hassan *et al*., 2006).

The importance of network measurement lies in the observation and understanding of networks. There are various studies focusing on this topic, both in passive and active modes of measurement.

Corresponding Author: Mustafa M.H. Ibrahim, InterNetWorks Research Lab, School of Computing, Universiti Utara Malaysia, 06010 UUM Sintok, Kedah, Malaysia

Yang *et al*. (2011) focused on high speed real time passive HTTP traffic performance measurement, they proposed architecture to real-time measure HTTP performance passively from traffic and they considered benchmark to validate the accuracy as future work.

Kim *et al*. (2009) focused on Internet traffic classification, this study critically revisits traffic classification by conducting a thorough evaluation of three classification approaches, based on transport layer ports, host behavior and flow features. In sight of this study was the effectiveness of port-based classification in identifying legacy applications is still impressive, further strengthened by using of packet size and TCP flag information.

Iliofotou (2009) proposed a graph-based representation of internet traffic which captures the network wide interactions of applications. They worked on TDGs in depth and they provide a graph that summarizes network-wide behavior of applications in order to classify network traffic.

Cao *et al*. (2010) focused on analyze three locality-awareness policies for Bit Torrent-like system: Tracker locality, choker locality and picker locality. Based on a Hsp, analyze how much network load saving can be expected for these locality policies, as well as their impact to the downloading efficiency of the system, as a result they become with that all locality policies can significantly reduce the average distance of both downloading and streaming scenarios.

Wang *et al*. (2011) focused on Internet traffic on the Tsinghua University campus, they analyzed the geographical origins of incoming flows and the result reveals that USA, Japan and Korea are the most important source countries of internet traffic. They also find that 74% of the international traffic volume is coming from only 5 countries, where in USA ranks first. There are only 7.3% of international http active source hosts in Japan; However, it provides 27.2% of http traffic and results show that the major has a stronger influence of users' average online time, while occupation has a stronger influence on users' average international traffic volume.

Augustin and Mellouk (2011) authors studied 20 popular web applications that are representative of 12 application types. As a result of this study authors propose a preliminary draft for a classification of Web applications according to the three traffic features traffic intensity is the total volume of traffic exchanged by the application during the measurement period, traffic symmetry measures the ratio between upstream and downstream traffic and traffic shape describes the general aspect of the traffic pattern. However, the study presented in this study is only a first step towards a precise classification of the plethora of HTTP applications available on today's Web.

Pries *et al*. (2009) focused on home users at a broadband wireless access service provider in order to reflect only home user traffic characteristics. They present the results of these measurements, showing daily traffic fluctuations, flow statistics as well as application distributions. The results show a difference to backbone traffic characteristics. Furthermore, they observed a shift from web and Peer-to-Peer (P2P) file sharing traffic to streaming applications.

Ihm and Pai (2011) focused on analyzing real web traffic during five years (2006-2010) from a globally-distributed proxy system, which captures the browsing behavior of over 70,000 daily users from 187 countries. Using this data set, they examine major changes in Web traffic characteristics that occurred during this period. They presenta new web page analysis algorithm that is better suited for modern webpage interactions by grouping requests into streams and exploiting the structure of the pages. Finally, they investigate the redundancy of this traffic, usingboth traditional object-level caching as well as content-basedapproaches.

In this study, we will focus on characterizing the Internet traffic based on users' preferences in Universiti Utara Malaysia main campus as a case study. Then, the result will be used to generate a guidance and suggestion to benefit all higher learning institutions. UUM provides Internet access for over 28,000 students and 6,000 staff members. Moreover, the campus is linked to the Internet through TM-ISP Internet provider. Every college in UUM is incorporated with Distributed multilayer switches which connect to a couple of core switches in the computer center. The computer center in UUM comprises varying high efficiency network devices including servers, firewall, controllers, multilayer switches and routers.

2. MATERIALS AND METHODS

Prior to initiating a data collection process, there are two main aspects that have to be taken into consideration, network topology and devices. It is imperative to determine the most suitable place for collection of packets in order to avoid capturing irrelevant packets. To conduct this research rigorously, we follow Jain and Hassan (2004) steps as shown in **Fig. 1**.

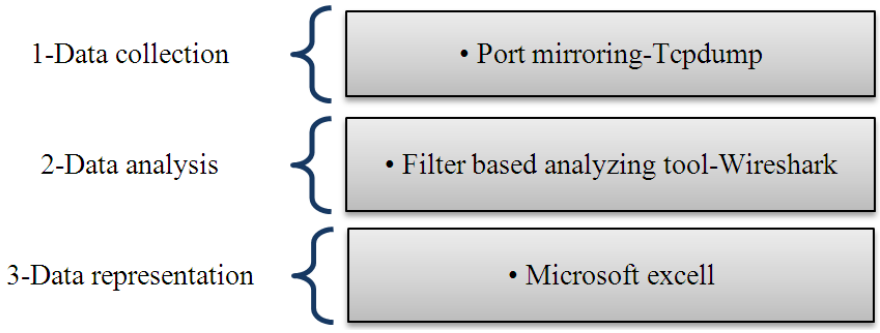

Fig. 1. Methodology (Jain and Hassan, 2004)

Table 1. Details of the captured data

Day	Date	Total Files Size	From To
Sunday	15/4/2012	37.8 Giga-Byte	10:00 until 11:00 AM
Monday	16/4/2012	36.3 Giga-Byte	10:00 until 11:00 AM
Tuesday	17/4/2012	41 Giga-Byte	10:00 until 11:00 AM
Wednesday	18/4/2012	39.1 Giga-Byte	10:00 until 11:00 AM
Thursday	19/4/2012	43 Giga-Byte	10:00 until 11:00 AM

2.1. Data Collection

Lenovo G560 was used as a capturing device and was connected to the main distribution Cisco's catalyst 1750 is the Internet switch in UUM computer center supporting mirror porting. This switch is connected to two main multilayer switches (Cisco's catalysts 6509) by fiber-optic data-transfer cables. Internet switch port Gigabit-Ethernet1/0/24 was mirrored to 1/0/11 port, this port connected to the bandwidth manager which is connected to the firewall in the computer center.

All data collected from the network is raw data divided and saved in the form of (Cap) extension files, libpcap library was developed using C programming language designed to convert network interface card NIC in promiscuous mode, that's guaranteed all packets arrived to the interface will be captured. Capturing device use Linux Ubuntu 12.04 LTS 64 Bit operating system and Tcpdump tool to capture mirrored packets. Total file size of the captured packets were 197.2 Giga-Byte divided as follows **Table 1**.

2.2. Data Analyzing

Wiresharkis the most well-known open-source network data analyzer. Wireshark is a steady and valuable component for all network toolkits (Orebaugh *et al*., 2007). Wireshark was selected for the packets' analyzing process by isolating HTTP data then isolate all HTTP request data and finally calculate all websites' statistics. GeoIP Domain Name and Geo Lite Country databases were implemented in the Wireshark program to specify each country, domain name and sub-domain name for each site.

2.3. Data Representation

Microsoft Excel program is used to present the data. This program can deal with mathematical problems, tables and presents graphical figures. Main steps is to transfer all websites' statistics, that's taken from different stages of analyzing' process to Excel spreadsheets, rows and columns have specified in ways that allow drawing figures, which exactly reflect the packet's status.

3. RESULTS

This study reports our analysis of the packet header traffic based on the application categories and signatures (domains) as shown in **Table 2**.

Following the analysis of the entire HTTP request packets process by UUM distribution switch and after specifying the websites traffic, it is clearly that social networking sites received the highest packets request percentage compared to other sites with 42% of the total request packets for all the five days within an hour. It is assumed that it is owing to the social networking sites' provision of video and URL sharing devices. This is followed by search engines with 19% which are indispensable to staff and student alike and E-commerce with 9%. **Figure 2** depicts the most preferred categories and their percentage.

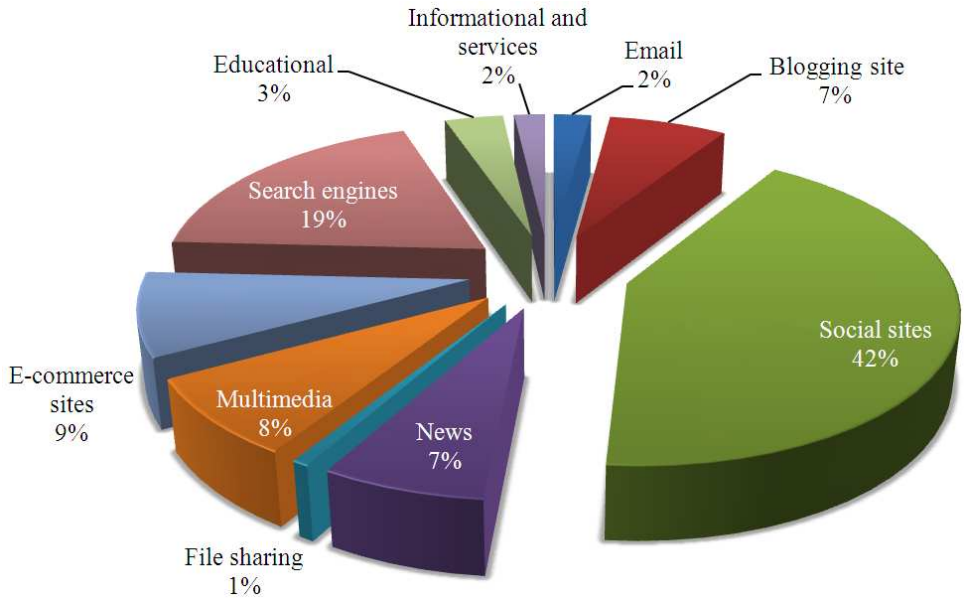

Fig. 2. Category of users' preferences

Table 2. Application category and signature

Category	Signatures
Social networks	Facebook; Twitter; Tagged; Digg;
Search engines	Google, Yahoo,bing
E-commerce and tech support	Airasia, Microsoft, Adobe, Mozilla
Education	
Multimedia	YouTube, Drectvid, Metacafe
News	Bbc, Wordpress, Thestar
Blogging	Blogger, Blogspot, beautifulnara
E-mail	Hotmail, Yahoo, Gmail
Informational and services	UUM, UTM,USM
File sharing	Mediafire, 4shared

The following will provide insight on the distributions of social sites, search engine, multimedia, educational site and blog because they are most related to teaching and learning activities using web 2.0 tools.

3.1. Social Preferences

One of the advantages of social-networking sites is the ability to connect users from all over the world, whether using text, audio and video sharing these are the main reasons of the widespread usage of social-networking sites. By analyzing social site's data for all the days, it became clear that Facebook took the highest request compared to the rest of social sites. This is a normal, since Facebook is a global site and it's one of the first sites that established the concept of sharing between users.

Figure 3 demonstrate the percentage of each social website for all days. Facebook took 94% of the social networking sites packets, while Tagged gained 3% and Twitter has 2%; MySpace and Digg took 1%.

3.2. Search Engines Preferences

Figure 4 illustrates the statistics for each search engine request packets. Google took the highest rate 97%, bing took 2 and Yahoo took 1%.

3.3. Multimedia Preferences

Some research indicates YouTube as a site for entertainment, others classify YouTube as a video sharing site and lie under this category, other researchers considered YouTube as a social-networking site, in our study; we list YouTube under multimedia category. **Figure 5** demonstrates the statistics of each Multimedia website. YouTube took the highest rate with 74% then radioactive.com broadcasting radio stations through the Internet it took 24% of multimedia packets. Directdive.com and metacafe.com are sharing video sites took they took 2%.

3.4. Blogs Preferences

In 2007, BlogSpot and Blogger considered as most famous blogging sites worldwide because of high participant's number of these sites, after analyzing the data, BlogSpot.com took the highest rate 77% of packets then blogger.com with 12% and finally beautifulnara.com gained 11%. **Figure 6** presents the percentage of each website.

3.5. Educational Preferences

Figure 7 illustrates the statistic of each Educational websites request packets.uum.edu.my took the highest rate 92%, it contains all service that fall under this domain then utm my took 4%.

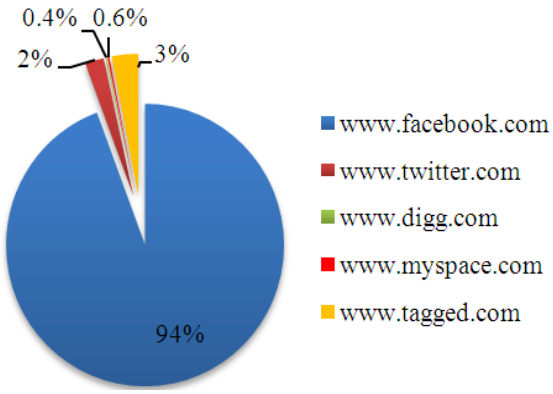

Fig. 3. Social networking sites distribution

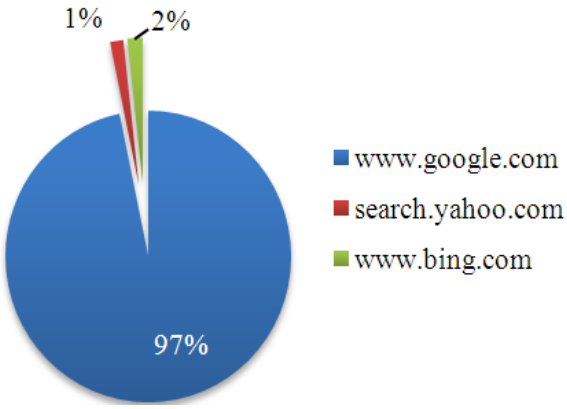

Fig. 4. Search engines websites packets rates

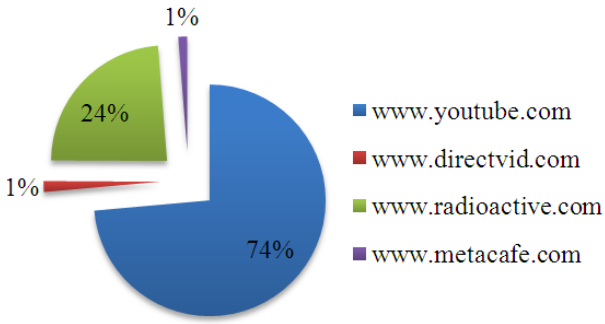

Fig. 5. Multimedia websites packets rates

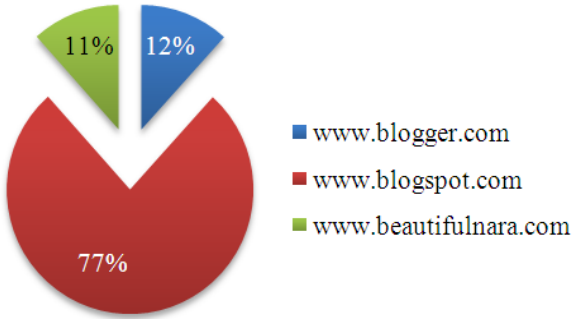

Fig. 6. Blogging websites distribution

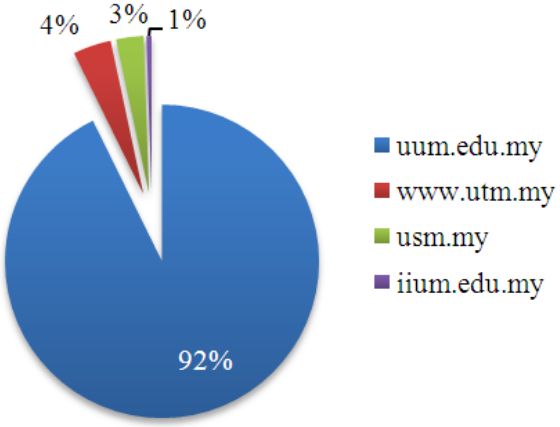

Fig. 7. Educational engines websites packets rates

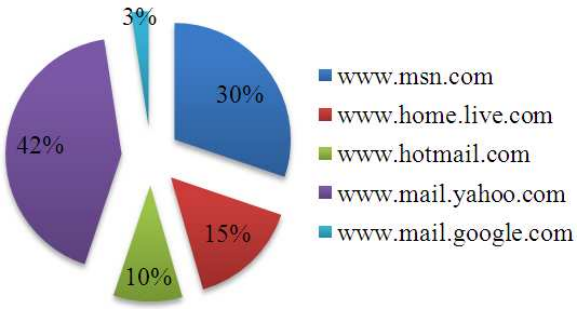

Fig. 8. E-mail websites packets rates

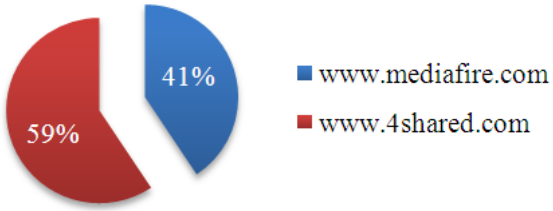

Fig. 9. File sharing websites packets

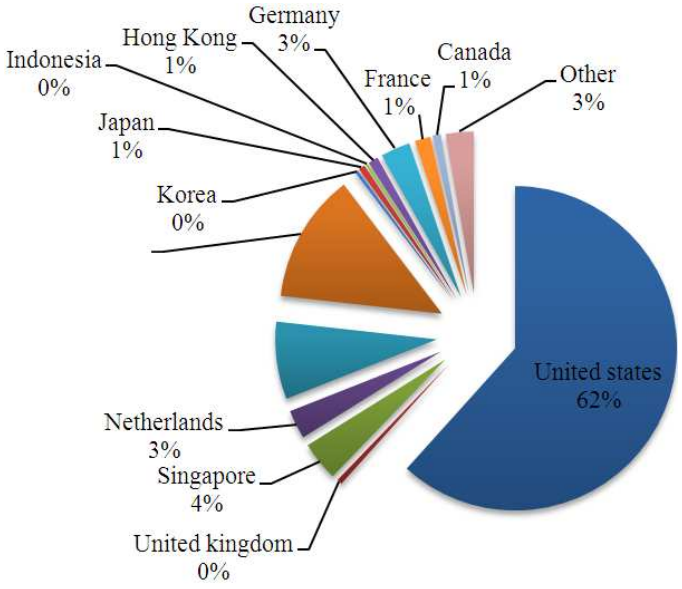

Fig. 10. Packets distributions over countries

3.6. E-mail Preferences

Figure 8 demonstrates the statistics of each Email website from 10:00 to 11:00 AM daily, yahoo mail and messenger took the highest rate 42%. Msn. comand msn messenger took 30%, other websites statistic shown in **Fig. 8**.

3.7. File Shearing Preferences

Figure 9 illustrates the statistics for each file sharing website.mediafire.com and 4shared.com are well-known domain in Internet world for sharing file. 4shared.com took the highest packets rate 59% and mediafire.com took 41%.

3.8. Countries Traffic Distributions

After completing the process of identifying Internet user's preferences, in UUM campus, starts a new process which is to specify and determine the direction of packets, depending on GeoLite database that implemented to Wireshark software. However, this process was based on analyzing HTTP packets, whether HTTP request or HTTP responds packets.

Figure 10 illustrates the packets destination of several countries. United State of America has the highest packets and this situation is typical, because main servers of most websites that has high packets rate are in USA, such as google.com, facebook.com, Twitter.com.

4. DISCUSSION

With regards to the first objective, the network usage in UUM registered high rates during all the five days, with the highest usage at 23730 PPS and the lowest at 21849 PPS. These statistics' variation depends on changes in time. The results revealed that several packets were lost during transmission which required the re-transmission of the corrupted packets and resulting in the increase of transmitted packets rates in the network.

Majority of the packets were of short-length leading to exhaustion of network devices during transmission. Further studies are called for to clarify this area of finding.

The findings also revealed that the used rate of TCP and HTTP protocols was high, characterized by many errors in the TCP protocol including out-of-order segment, Zero window packets, Duplicate acknowledgement among others.

As for the second objective, through the analysis of the users' behavior, the findings revealed that the most frequently visited sites were that of social networking sites and video streaming implying the need to limit such access during working or studying hours.

5. CONCLUSION

Through the analysis of the users' preferences, the findings revealed that the most frequently visited sites

were that of social networking sites and video streaming implying the need to limit such access during working or studying hours.

Generally speaking, users' applications within the network must be analyzed to confine the highly-used applications like streaming applications and determine its effects of the network.

Designing a new system similar to internal social networking sites may allow staff and students to share information and communicate effectively in contrast to global social networking sites. Some policies such as firewall blocks in social networking sites may be modified during working or studying hours. Applying strict policies on bandwidth control manage to limit the use of non-beneficial Internet applications in the scientific field and monitoring the network to detect new web-applications behavior.

The findings of the present study may also be used as reference for comparative network performance studies in the future and this may be helpful for network administrators in their plans to improve network efficiency by determining which Internet resources negatively affect the network. The findings also provide valuable insight to network engineers in their attempts to design and improve better network performance in light of the users' behavior and requirements.

6. ACKNOWLEDGEMENT

Researcher would like to thank and appreciate UUM Computer Center for the immeasurable support to conduct this research.

7. REFERENCES

Augustin, B. and A. Mellouk, 2011. On Traffic Patterns of HTTP Applications. IEEE Commun. Foci. J., 11: 2-4.

Cao, Y., B. Liu and Y. Xue, 2010. Locality analysis of bittorrent-like peer-to-peer systems. Proceedings of the 7th IEEE Consumer Communications and Networking Conference, Jan. 9-12, IEEE Xplore Press, Las Vegas, NV., pp: 1-5. DOI: 10.1109/CCNC.2010.5421708

Hassan, H., J.M. Garcia and C. Bockstal, 2006. Modeling internet traffic: Performance limits. Proceedings of the International Conference on Internet Surveillance and Protection, Aug. 26-29, IEEE Xplore Press, Cote d'Azur, pp: 4-7. DOI: 10.1109/ICISP.2006.20

Ihm, S. and V.S. Pai, 2011. Towards understanding modern web traffic. Proceedings of the ACM SIGCOMM Conference on Internet Measurement Conference, Nov. 02-04, ACM Press, New York, pp: 295-312. DOI: 10.1145/2068816.2068845

Iliofotou, M., 2009. Exploring graph-based network traffic monitoring. Proceedings of the 28th IEEE International Conference on Computer Communications Workshops, Apr. 19-25, IEEE Xploe Press, Rio de Janeiro, pp: 757-758. DOI: 10.1109/INFCOMW.2009.5072143

Jain, R. and M. Hassan, 2004. High Performance TCP/IP Networking: Concepts, Issues and Solutions.1st Edn., Prentice Hall, ISBN-10: 0131272578, pp: 383.

Kim, H., K. Claffy and M. Fomenkov, 2009. Internet traffic classification demystified: Myths, caveats and the best practices. Proceedings of the International Conference and Workshop on Emerging Trends in Technology, (TT' 09), pp: 891-893.

Orebaugh, A., G. Ramirez and J. Burke, 2007. Wireshark and Ethereal: Network Protocol Analyzer Toolkit. 1st Edn., Syngress Press, ISBN-10: 1597490733, pp: 540.

Pries, R., F. Wamser, D. Staehle, K. Heck and P. Tran-Gia, 2009. On traffic characteristics of a broadband wireless internet access. Proceedings of the Next Generation Internet Networks, Jul. 1-3, IEEE Xplore Press, Aveiro, pp: 60-62. DOI: 10.1109/NGI.2009.5175772

Rogers, P.C., S.W. Liddle, P. Chan, A. Doxey and B. Isom, 2007. A Web 2.0 learning platform: Harnessing collective intelligence. Turkish Online J. Dis. Educ., 8: 16-33.

Sheely, S., 2006. Persistent technologies: Why can't we stop lecturing online. Proceedings of the 23rd Annual Ascilite Conference: Who's Learning? Whose Technology? (WT' 06), University of Sydney, pp: 769-774.

Sims, R., 2006. Online distance education: New ways of learning; new modes of teaching? Dis. Educ., 27: 3-5.

Wang, J.H., C. An and J. Yang, 2011. A study of traffic, user behavior and pricing policies in a large campus network. Comput. Commun., 34: 1922-1931. DOI: 10.1016/j.comcom.2011.05.009

Yang, X., X. Chen and Y. Jin, 2011. A high-speed real-time HTTP performance measurement architecture based on network processor. Proceedings of the International Conference on ICT Convergence, Sept. 28-30, IEEE Xplore Press, Seoul, pp: 744-745. DOI: 10.1109/ICTC.2011.6082715