American Journal of Applied Sciences

Literature

Online Traffic Measurement and Analysis in Big Data: Comparative Research Review

[1]Lena T. Ibrahim, [1]Rosilah Hassan, [1]Kamsuriah Ahmad, [1]Asrul Nizam Asat and [2]Halizah Omar

[1]*Research Center for Software Technology and Management (SOFTAM), Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia*
[2]*Pusat Citra UKM, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia*

Article history
Received: 16-11-2015
Revised: 23-04-2016
Accepted: 27-04-2016

Corresponding Author:
Lena T. Ibrahim
Research Center for Software Technology and Management (SOFTAM), Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia
Email: lena_alkazaraji@yahoo.com

Abstract: The Internet traffic measurement and analysis is important to avoid many problems of data transferring via online networks such as data losing and slow data transferring. The Internet traffic measuring and analysis could be effective to avoid the data transferring challenges. The Internet traffic measuring and analysis flexibility is important due to many reasons such as dynamicity of transferred data such as size and format, the data transferring protocols and the dynamicity of measure the traffic based on the networks available resources depend on the transferred data characteristics. The main objective of this paper is to review the most flexible Internet traffic measuring and analysis tools that could be adopted to handle the dynamicity of data transferring characteristics. The significance results show that the Hadoop/MapReduce tool has many advantages over other traffic measuring and analysis tools. The Hadoop/MapReduce features are easy to be modified based on various selections of Internet traffic measuring, the Hadoop/MapReduce is compatible with various format of data transferring such as texts, videos and images and the Hadoop/MapReduce can analyze the better ways of data transferring depend on many transferring protocols such as Transmission Control Protocol (TCP) and User Datagram Protocol (UDP).

Keywords: Traffic Measurement, Traffic Analysis, Hadoop/MapReduce, Data Transferring, Flexibility

Introduction

Network monitoring and measurement have gained greater importance in modern day complicated network. Previously, network administrator monitor only a limited number of network devices or computers whose number range less than a hundred (John, 2008). Network bandwidth during those days was only at 10 or 100 Megabit per second; however, the current online network bandwidth is higher as it is used to gather large volumes of data and information amongst people (Parekh and Patel, 2015) and the efficient functioning of the network of spam interventions depends on the routing protocol (John, 2008). For example, the social networks application such as Facebook transfers millions of data daily in images and videos format. Therefore, one of the major issues of networks monitoring is the speed and performance of data transfer. The availability of networks or online traffic may lead to the delay of transfer time for data. In view of this problem, researchers have developed tools to measure and analyze the online traffic in order to analysis and avoid problems that may slow down data transfer processes. In the case of network failure, monitoring are needed to automatically discover, separate and correct network breakdowns and most probably make up for the failure. Generally, the agents are required to send warning to the administrators to fix the problems. When the network is stable, the administrator's responsibility remains to regularly monitor in case there is inside or outside threat to the network. In addition, agents also have to watch regularly to figure out how overloaded the network device(s) are. The work of this agent was recorded in the log files for future reference and troubleshooting. Information or log about the use of network can be utilized to enable network to function effectively to present a failure and for future improvement (So-In, 2014). To handle and perform network monitoring and analysis, different types of method can be applied such as Hadoop/MapReduce and network flow monitoring and analysis tools (Hasib and Schormans, 2003).

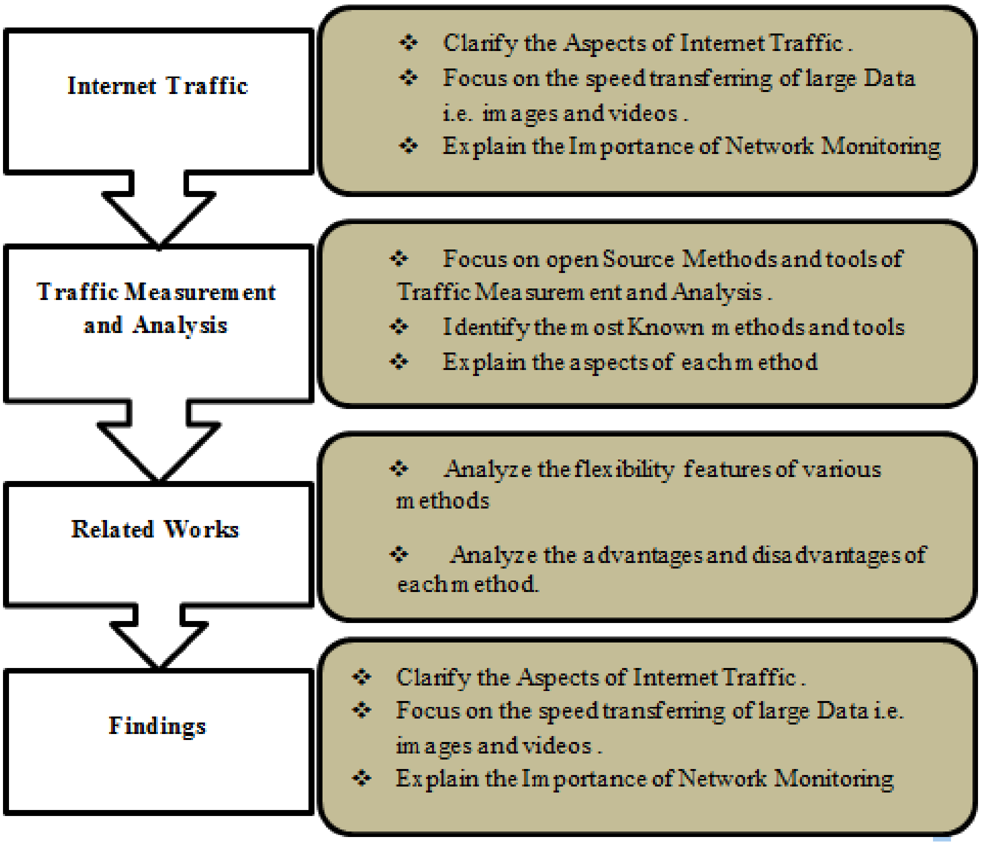
Internet Traffic
Clarify the Aspects of Internet Traffic.
Focus on the speed transferring of large Data i.e. images and videos.
Explain the Importance of Network Monitoring
Traffic Measurement and Analysis
Focus on open Source Methods and tools of Traffic Measurement and Analysis.
Identify the most Known methods and tools
Explain the aspects of each method
Related Works
Analyze the flexibility features of various methods
Analyze the advantages and disadvantages of each method.
Findings
Clarify the Aspects of Internet Traffic.
Focus on the speed transferring of large Data i.e. images and videos.
Explain the Importance of Network Monitoring

Fig. 1. Paper structure

One of the challenges that face the methods of Internet traffic measuring and analyzing is the flexibility weakness, which can be characterized into the following:

- Calculating the network usage as two main fixed classification i.e., heavy networks or free network (Dittrich and Quiané-Ruiz, 2012)
- The number of data segmentation is fixed in the case of heavy network usage (Dittrich and Quiané-Ruiz, 2012)
- The traffic analysis processed based on TCP as main transfer protocol (Liu *et al*., 2014)

Through various activities, the Internet traffic measurement and analysis processes flexibility could be enhanced which includes calculating the network usage based on various efficient classifications such as low, medium and heavy traffic. The number of data segmentations can be divided dynamically based on the current classification of network usage, the transferred data characteristics and the data size. In addition, for the purpose of improving the traffic analysis and data transfer processes, the use of TCP and UDP protocol may be adopted.

Figure 1 given illustrates the structure of this paper. Section 2 reviews the definitions, aspects and tools of Internet traffic measurement and analysis while section 3 discusses the related works of various tools of measuring and analyzing Internet traffic. Section 4 presents the findings of related works discussions and section 5 provides the recommendation of this paper based on the findings. Towards the end, the results of the paper are summarized.

Literature Review

The literature review discusses the theoretical and practical aspects of various Internet traffic measurement and analysis tools. It seeks to simplify and compare the advantages and disadvantages of different tools keeping in view the main objectivities of this paper.

Importance of Network Monitoring

Over the past decade or so, the use of Internet between people has rapidly increased which act as an important vehicle of communication. The online data transfer and file sharing have also increased rapidly through many Internet applications such as emails, social networks; multimedia or video services and chats (Gebert *et al*., 2012). Thus, the Internet usage also has undergone heavy changes. Similarly, the heavy usage of

Internet leads to increase in the Internet network traffic (Pries *et al*., 2009).

According to Lee and Lee (2013), the Internet traffic can leads to experience unexpected delay in data transfer and sharing.

There are several advantages when considering network monitoring. Some of these advantages consist of providing a clear view of network availability and performance. This includes reducing the cost associated with the asset utilization which as a result reduces the risk by providing a secure network in accordance with the standard network guidelines. In addition, network monitoring also help to achieve service level agreements and documenting progressive performance with reports. Thus, it become crucial to determine effective mechanisms to avoid loss and delay in Internet traffic based on the use of different network protocols and routing algorithms. Those approaches and protocols are mainly aim at measuring and analyzing the network bottlenecks to efficiently manage the online data transfer processes (Liu *et al*., 2014).

Internet Traffic

Internet communication connects millions of Internet users across the world (Awduche *et al*., 2002; Fortz and Thorup, 2000). The Internet communication network encapsulates a massive optical fibre infrastructure, wireless connections and a chain of copper wires which connect large number of end hosts utilizing packet switches that range from conventional web servers, personal computers, to mobile phones and other smaller devices usually used at homes, in cars and in the day-to-day routine. As a massive infrastructure system, Internet also supports an array of applications. Those applications include World Wide Web, emails, file sharing, telephony, radio, video games and other commercial services.

It is a crucial question to address as how to articulate or describe Internet traffic efficiently. Internet traffic characterizes as to when, who and how traffic may be investigated. As compared to a larger company, the behaviors of traffic network differ considerably from that of traffic behaviors in a smaller company. It may also be noted that with new applications, characteristics and behavior of traffic can change such as in networks that are of new type.

A survey conducted measured that about 70 to 75% of the traffic was comprised of web traffic (Fortz and Thorup, 2002). From there onward, we have witnessed significant increase in the total traffic volumes web traffic still occupies larger share of many networks (Yuan, 2001; Awduche *et al*., 1999; Bhattacharyya *et al*., 2001). However, the sharing of file often dominates the traffic application (Ahuja *et al*., 2015; Poppe *et al*., 2000). In addition, TV and video distribution over IP have become widely spread and produce increasing traffic volumes. Therefore, the Internet traffic is dependent on the network resources and the transferred data through these resources i.e., transfer speed at same time. For instance, the transferred volume of data may lead to Internet traffic on low-resourced networks while in the large network resource, the congested traffic may not occur.

Internet Traffic Measurement and Analysis

Currently, the Internet appears to have emerged as the key applications utilized widely for the purpose of personal, official and commercial communication. One of the major contributing factors to the ongoing phenomenal growth and expansion of Internet is its incredible qualities such as versatility and flexibility. The versatility and flexibility of the Internet can be gauged from the fact that we can connect any electronic or digital device to the Internet now that might range from conventional desktop/personal computers to supercomputers or larger servers covering many kinds of wires devices such as hand phones, sensors, etc. In addition, we can also witness dramatic changes and drastic transformations in the usage of Internet, much different from that of the very earlier use ever since 1969. A project in 1969 enabled a small number of terminals to facilitate in a limited range of far-off operations (John, 2008). Currently, the Internet can serve as the key data transmitter and deliverer for a larger range of protocols, enhancing opportunities of exchange of not only textual data, but also that of voice, audios, videos and several other different modes of digital media connecting millions of users from across the world (John, 2008; Arlitt and Williamson, 2005). We may also note that one of the impediments is that the speedy growth does not leave sufficient time and resources to assimilate measurement and analysis possibilities into Internet infrastructure, applications and protocols. We can understand network infrastructure and individual protocols when we test them in isolated lab environments and in network simulations. But, particularly, global-scale hostile Internet environment is generally not clear as a large number of Internet applications interact (Arlitt and Williamson, 2005). The challenges in developing full understanding is further multiplied by the fact that the “shape” of the Internet was not planned in advance, where diverse networks of autonomous organizations have been connected to the main Internet. Hence we witness that the protocols and applications of Internet not only transform and evolve with time, but they also travel and shift across geographical territories. Simultaneously, one of the growing concerns is that the enhanced bandwidth and increasing Internet users has also ensued its misuse and inconsistent behavior (Arlitt and Williamson, 2005;

Brownlee and Claffy, 2002) which call for analysis so that effective and suitable counter strategies against the misuse may be auctioned. Speaking from a broader perspective, this signifies that even though, we may describe Internet as the most vital modern communication vehicle, there are still crucial questions that lurk as to why and how Internet functions. As we find, majority of network operators fail to supply reasonably solid answers as to figure out the traffic that runs on their network. It is so because the Peer to Peer (P2P) applications of file sharing in the newly surfaced network can be disguising.

This may be admitted that the Internet using community does recognize the vibrant character and behaviour of contemporary Internet traffic to carry out research for further development of the Internet. To develop thorough understanding of the modern Internet requires a measurement of Internet traffic, ideally speaking on highly aggregated links. It is critical to point out that the accurate measurement of Internet traffic involves an array of complex tasks and challenges. However, it may also be added that as soon as we overcome the practical, technical or legal complications of the Internet traffic, we can then potentially achieve the solidity of protocols, infrastructure mechanisms and performance systems (Brownlee and Claffy, 2004).

The past couple of decades have witnessed the development of a larger range of tools to monitor the Internet traffic. One amongst such tools is known as Tcpdump (Brownlee and Claffy, 2004), which is used to capture and analyze packet traces with libpcap. In addition, CoralReef which has been developed by Center for Applied Internet Data Analysis (CAIDA), which provides a flexible traffic capturing facility, analysis and report functions. Another tool that can help measure Internet traffic is known as Snort (Roesch, 1999). Snort is an open source signature-based tool designed for instruction and detection giving support in real-time analysis. On the other hand has been designed to bolster the cluster environment. Other tools used for the same purpose includes Tstat and L7 Filter (Finamore *et al.*, 2010). Both tools are a passive analysis tool which uses Tcptrace and can function for several analysis capabilities for TCP performance metrics, application classification and VoIP characteristics. On the other hand, NetFlow is another recognized flow monitoring format designed for observation of Internet traffic with the help of switches and routers. Finally, Hadoop/MapReduce (Lee *et al.*, 2011) is another application that has been designed for analyzing web, larger texts and log files and so on. It may be critical to point out that although there exist larger number of tool sued for monitoring and measuring Internet traffic; majority of the tools mentioned above usually operate on a single server environment.

Tools of Online Traffic Measurement and Analysis

This section presents the overview of ten known online traffic measurement and analysis tools to clarify the aims and architecture of these tools. There are many traffic measurement and analysis tools, but the most widely used are 7 tools are the following: (1) Hadoop/MapReduce, (2) Tcpdump, (3) CoralReef, (4) Snort, (5) Tstat, (6) NetFlow, (7) L7 Filter, (8) Pandora FMS, (9) Microsoft network Monitor and (10) Angry IP scanner.

Hadoop/MapReduce

Hadoop/MapReduce can be described as a software framework that is usually used for easy applications such that of writing. It is a reliable tool that processes and places larger clusters of commodity hardware and process a large amount of data in parallel (Liu *et al.*, 2014). The function of MapReduce is generally to divide the input data into independent sets, performing in parallel manner to process several map tasks. In general, the input and output data may be stored using filing system. The framework performs several functions such as it monitors and schedules tasks and performs the tasks of failure. Usually, the storage nodes and computer are similar which means that the MapReduce framework and the Hadoop Distributed File System (HDFS) operate on similar sets of nodes (Quick and Choo, 2013). This configuration enables the framework to schedule tasks effectively on the nodes where data already exists. This results in resulting in a rather high aggregate bandwidth across the cluster. This framework consists of a single master Job Tracker and one slave Task Tracker per cluster node. The master monitors and executes the failed tasks. In addition, its job is to schedule the jobs' component tasks on the slaves. As directed by the master, the slaves execute the tasks minimally, with the help of performing appropriate interfaces or abstract classes, this application identify the input/output locations, map and reduce functions. Subsequently, Hadoop job client submits the job and configuration to the Job Tracker, which distributes the software/configuration to the slaves, scheduling tasks and monitoring, providing status and diagnostic information to the job-client (Fusco and Deri, 2010). Although, we implement the Hadoop framework in Java TM; however, MapReduce applications are not written in Java. Likewise, two utilities can be used to implement MapReduce One, Hadoop Streaming utility allows users to create and run jobs with any executable as the mapper and/or the reducer. Two, another utility is that Hadoop Pipes is a SWIG- compatible C++ API Application Programming Interface (API) can be used to implement MapReduce applications.

Tcpdump

Tcpdump can function in most Unix-like operating systems Linux, this tool is used to record network traffic.

By saving the traffic in diverse formats, this tool can help capture packets using wide range of user-specified criteria. Running under the command line, Tcpdump is a usually used to analyzes packets. The user is allowed to display TCP/IP and other packets being transmitted or received over a network to which the computer is attached (Ramakrishnan and Rodrigues, 2001).

In those systems, Tcpdump runs the libpcap library to capture packets in the systems. Another function is that of printing the contents of network packets. Tcpdump has the capacity to read packets from a network interface card as well as from a packet file that has been created previously. It also offers the possibility to display and intercept communication from one user to another or from one computer to another (Abrahamsson, 2008).

To put hold on the number of packet detected by Tcpdump, the Internet user may by choice, apply One of the advantages of this application is that this may turn the output comparatively more usable on those networks that run a high volume of traffic, there is also a possibility to drop down the privileges of a specific user once capturing mechanism has been put on action. In other systems such as Unix-like operating systems, the packet capturing mechanism can be configured to allow non-privileged users to use it; if that is done, super-user privileges are not required (Gunnar *et al*., 2005).

As a result, it is recommended that we do some analysis to get output of raw packets. However, there arises the problem of the incompatibility of the trace format such as “Microsoft Network Monitor” that cannot read the trace file from “Tcpdump”. On account of the performance issue, “Tcpdump” functions only as the traffic-capturing tool and “Tcpdump” can just capture the packets and saves them in a raw file. It can record the time elapsed, trip times, the segments and bytes delivered, the transmissions received and the window advertisement (Abrahamsson, 2008; Sridharan *et al*., 2003).

CoralReef

It is a comprehensive package used for different program languages. It can particularly use device drivers, written applications and libraries. The applications of CoralReef are mostly of two categories. First are those names that begin with CoralReef that relate to raw packet data and those names that run on aggregated flow data (Keys *et al*., 2001).

Custom Coral drivers, the libpcap library for commodity network interfaces and trace files generated by CoralReef trace, Tcpdump, or other software are normally the sources of raw data. All applications of the CoralReef take a common set of command line and configuration options. Applications that occur regularly carry a common syntax to specify interval size. The main utilities of CoralReef are:

- CoralReef trace captures network traffic to a CoralReef trace file
- CoralReef info reports hardware and link configuration details of a trace file
- CoralReef encode the IP addresses in a CoralReef file to protect privacy
- CoralReef hits reports packet and byte counts by IP length and protocol, port summary matrices for TCP and UDP, fragment counts by protocol, packet length histograms for the entire trace and for a list of applications and the top 10 source and destination port numbers seen for TCP and UDP traffic
- CoralReef flow at regular time intervals, aggregates packet data into flows by source and destination IP addresses protocol and source and destination ports (Keys *et al*., 2001; Brownlee, 1997)

Snort

A free and open network tool, snort is used for instruction, detection and prevention system designed by (Roesch, 1999). It can perform real-time traffic analysis and packet logging on Internet Protocol (IP) networks. It can perform protocols like traffic analysis, search of content and match of content. The program can also be used to probe attacks, common gateway interface and operating system fingerprinting attempts (Mehra, 2012). In three main modes can snort be configured: Sniffer, packet logger and network intrusion detection (Rafeeq, 2003; Mehra, 2012). The function of sniffer modes is to read the network packets and show them on the console in a continuous stream. Likewise, the packet logger mode logs the network packets to the disk. Finally, network intrusion detection mode can be described as the most complex mode. It monitors network traffic and analyzes it against a rule set defined by the user. Later based on identifying, it performs a specific action. Multiple components constitute snort. These components function in coordination to identify particular attacks.

NetFlow

With the help of Netflow switching feature, Cisco routers can produce network flow records. Additionally, it can be exported in either User Datagram Protocol (UDP) or Stream Control Transmission Protocol (SCTP) packets to NetFlow collectors (So-In, 2009). We can define NetFlow as a version number. Version 5 is rather commonly used one while version 9 is an Internet Engineering Task Force (IETF). Standard for Internet Protocol Flow Information eXport (IPFIX), sequence number, timestamps for the flow start and finish time, number of bytes and packets observed in the flow, Internet Protocol (IP) headers (Source and destination IP addresses, Source and destination port numbers, IP protocol, Type of Service value), the union of all Transport Control Protocol (TCP) flags observed over the life of the flow.

We find that the network flow information can be very valuable not only to understand network behavior, but it can also help identify security holes. Moreover, with the help of this, correct decisions can be made on network planning. For example, to determine who originates or receives the traffic, source and destination addresses can be used for this purpose. From part information, the application utilizing or distributing can be made. Class of service examines the traffic priority. Explained that the packets and byte count show the amount of traffic. For calculation of packets and byte count per second, flow timestamps can be used.

NetFlow record is cached when traffic is first passed by Cisco router and sent to the NetFlow collector on the following conditions: First, for TCP traffic, when the TCP connection is terminated. Secondly, when the flow is inactive in a certain time (default is 15 sec) and thirdly when the active flow is long lived (30 min by default) and finally when the flow table is full. However, these timers can be reconfigured. Furthermore, general NetFlow collectors provide a traffic flow aggregation feature.

Once the flow records are exported, the router does not store those flows on account of performance reason. Hence, there is no retransmission mechanism with UDP transmission because of the loss of flow packets. In terms of router's CPU consumption, collecting NetFlow data can be rather expensive The NetFlow collector is placed just one hop from the router or directly connected. Furthermore, “Sampled NetFlow” feature is an option in order for router to look at the packet in every packet or randomly selecting interval. Aside from the above recommendations such as placing the NetFlow collector, the location is also subject to the position of reporting solution and the topology of the network. However, NetFlow is placed on the main website because the implementation of NetFlow from the remote branch is optimal.

L7 Filter

L7Filter is an open source project that was publicly released in 2003 when it was becoming apparent that port-based classification techniques were unreliable L7 Filter is an application-level classifier that was originally designed for use with Linux NetFilter to perform traffic shaping and accounting. L7 Filter compares packet payload against a series of pre-defined signatures (described using regular expressions) and identifies the application based on what signature, if any, is matched by the packet payload. L7 Filter includes signatures for many application protocols, including well-known applications such as TCP (Karagiannis *et al*., 2004; 2005). The most recent release of signatures was in May 2009; therefore, so recently released applications are unlikely to be supported by L7 Filter. Historically, the L7 Filter signatures have been popular within the traffic classification community. During the same period, researchers requiring a free deep packet inspection Dots Per Inch (DPI) tool to provide ground truth data for testing and evaluating classification techniques, found that L7 Filter was the only feasible option (Grajzer *et al*., 2012; Dong *et al*., 2013; Carela-Espanol *et al*., 2011).

Tstat

Tstat is an open-source traffic analysis tool which comprises of an application-level classification component, Founded on deep packet inspection (Finamore *et al*., 2011). Contrary to other tools, Tstat does not aim at classification as major goal. The software is employed for a broader analysis of Internet traffic; therefore, as compared to other classifiers, it is expected to support fewer application protocols. In several studies recently, Tstat has turned out to emerge recreantly in literature as source of ground truth (Finamore *et al*., 2011; Adami *et al*., 2012; Grimaudo *et al*., 2012).

Pandora FMS

Pandora Flexible Monitoring System (Pandora FMS) is a software developed using visual way to monitor computer networks based on status and performance of several parameters from different operating systems, servers, applications and hardware systems such as firewalls, proxies, databases, web servers or routers.

Pandora FMS is used to remotely monitor several known protocol with the use of agents on any operating system. Agents are daemons or services that can monitor any numeric parameter, Boolean status, string or numerical incremental data and/or condition. It can be developed in any language based on their operating development platform and should be able to communicate with the Pandora FMS Servers using available data transfer protocol such SSH, FTP and NFS by utilizing the XML. Pandora FMS was also used for network security hardware monitoring via the TCP/IP stack (Parekh and Patel, 2015).

Pandora FMS uses WMI protocol to gather and process Windows based information from sources. In order to gather those information, Pandora utilizes multiple servers each with own functionalities that is for network discovery, inventory collection, predicting complex user-defined network test, replicating multiple Pandora FMS sites and gathering SNMP (Parekh and Patel, 2015).

The setup of multiple servers is vital for Pandora FMS as it gathers all information from numerous sources which enable them to generate alarms for monitoring activities. The configuration setup modular which is dependable on size of network structure, as a single system with multiple servers is sufficient for small network while big systems acquired multiple individual servers. Gathered data within all servers are required to be inputted into a central Pandora database and it is possible to connect it from multiple Pandora servers with different functionalities.

Microsoft Network Monitor

Microsoft Network Monitor (Netmon) was originally developed by Raymond Patch and Microsoft LAN Manager Development team for troubleshooting applications problem on the network. This packet analyzer can be used to capture, view and analyze network data and decipher network protocols (Quick and Choo, 2013).

Due to the high cost of acquiring a hardware-based analyzer, the team had to share a single machine. While testing on reproducing a network bug with the hardware, the idea of Netmon was conceived. The first 4 bytes of the Netmon capture file format were used to validate the file. Netmon uses 'RTSS' values which was derived from the four initial team members that is Ray, Tom, Steve and Steve. The development environment was on OS/2 with no user interface and therefore a symbol was placed in the device driver where the packet buffers were kept so received data could be dumped in hex from within the kernel debugger (Microsoft.com, 2012).

As networks and e-mail were not encrypted at the time and due to high cost of hardware analyzer caused a lot of problem for Microsoft IT in monitoring user access as Netmon provides network engineers free access to traffic. Improvement on identification features were added into Netmon by adding an identification protocol named the Bloodhound-Oriented Network Entity (BONE) and a non-cryptographic password (Microsoft.com, 2012).

Network Monitor initial main purpose is collecting all data related for analyzing security and future forensic but not network traffic. Instead of gathering data on relevant packets or frames, Netmon gathers the host information (Parekh and Patel, 2015).

Angry IP Scanner

Angry IP Scanner (IPSCAN) is an open-source network tool developed for ease of use on the multi operating system environment. This network scanning tool is available freely and used frequently by system administrators, individual business users and networking students in various organizations across the world (Angryip, 2014).

Angry IP works by initially sending ping command to destination host checking whether the IP is alive. If the ping is successful, it will then resolve basic information such as the hostname and gather information on open ports, MAC address and other relevant information. Numerous plugins was developed by its supporter to gather additional data from the targeted host (Angryip, 2014).

The added plugins was also able to collect user based information such as the computer user or name, workgroup name and NetBIOS information. This information can be captured into standard file format such as TXT. It uses multithreaded approach where a separate scanning thread is created for each scanned IP address, in increasing its scanning speed (Gadge and Patil, 2008).

Angry IP scanner scans IP address for ports within alive hosts but once open ports is not detected on destination host, it will consider them as filtered (Parekh and Patel, 2015).

Comparisons Review of Internet Traffic Measuerment and Analysis Tools

Practical studies suggest that (Alcock and Nelson, 2013), Hadoop/MapReduce tool offers a variety of flexibility advantages for measurement and analysis processes of internet traffic. Adopting online network simulation, the researchers test Hadoop MapReduce. The most significant advantages include: Ability to write MapReduce programs in Java, a language which even many non-computer experts can manage to learn with adequate ability to account for powerful data-processing needs. In addition, it makes us capable of rapidly processing a huge amount of data at a time. Furthermore, contrary to expensive, specialized parallel-processing hardware, it can be effectively applied on large clusters of cheap commodity hardware as. Also, drawing on network capabilities such as networks speed and network usage, it can help update the data segmentation number. Finally, the other advantage is that it can transfer data using a variety of protocols that include TCP and UDP. It is also crucial to identify some of the limitations which Hadoop/MapReduce has. For instance, Procedural programming model entails code even for the very basic operation (projection, filtering). Another limitation is that Map Reduce nature is not specifically aimed to implement codes that have iterations or iterative behavior (Kadam and Dhore, 2010).

A study was conducted to analyze traffic measurement and analysis tools including NetFlow Drawing on dynamic environment of measurement and analysis processes, the study was mainly aimed to analyze the tools performance. The survey analysis focused on implementation flexibility that is one of the most important indicators. NetFlow tools have many advantages such as it can be integrated with various transferring protocols like IP/ICMP/UDP/TCP. It assists in real time data collection with various networks speed and it can work in different type of data such as images, audio and text files. Notwithstanding its advantages, NetFlow can have a number of drawbacks. It is not compatible not compatible with windows operating system. NetFlow can be weak in capturing network usage. For example, it has limited capacity to measure traffic and it cannot be used to manage data segmentation. Other limitations include that it cannot be scaled and it at times mismatch with size of transferred data and the network usage.

According to (So-In, 2009), L7 filter and Tstat approaches are open sources tools that can be applied in Linux operating systems. So-In (2009), Conducted practical comparisons between different traffic measurement and analysis tools that included L7 filter and Tstat approaches. Deploying a variety of network capabilities and different transferring protocols, comparisons were drawn to account for the ability of measuring and analyzing the internet traffic (Alcock and Nelson, 2013).

The testing dataset were comprised of numerous applications that included YouTube, Facebook, Twitter, FTP Data, Gtalk and iTunes Store. Results suggested that the L7 filter is the least performing approach of all the approaches tested. Contrary to this, Tstat approach proved comparatively better in terms of performance in the traffic measurement. Nevertheless, it can still be termed less efficient in the traffic analysis. Therefore, we suggest that both L7 filter as well as Tstat approaches may be recommended with a certain degree of caution while applying them in dynamic environment of internet traffic measurement and analysis (Alcock and Nelson, 2013).

According to (Abrahamsson, 2008), Compares between the Snort method and other traffic measurement and analysis tools. The comparisons involved many variables such as features flexible.

Customization, high speed network capability and operating system compatibility. The methodology of comparisons involved practical network simulation applying different measurement and analysis process of the given approaches. Result suggested that Snort demonstrated records medium flexibility performance in the features customization. Snorts have been recorded showing medium flexibility performance in high speed network capabilities whereas it is weak on the low-speed networks. Additionally, the snort can be used operating systems such as windows and Linux.

According to (Brownlee, 1997), one of the key the key disadvantages of CoralReef approach is that it can monitor traffic only that is observable to a network interface. To monitor a link between routers or on a switched network, it entails pointing traffic into added dedicated interfaces, which may either be standard interfaces read via libpcap, or special hardware accessed through Coral drivers. The needs of the hardware depend on the use of the monitoring of links and the amount of desired aggregation. The main constraint for straight forward packet traces is generally disk performance and capacity Likewise, memory and CPU speed are more important for flow collection and analysis, Hence, for the purpose of internet traffic measurement and analysis, the CoralReef required to be developed better the implementation of Tcpdump too involves the nearly same problem. The problem arises because the Tcpdump is mismatched with the trace format as "Microsoft Network Monitor" cannot read the trace file from "Tcpdump". Thus, rather than analysis processes, "Tcpdump" functions may be applied only as the traffic-capturing process. For the purpose of traffic analysis and data transferring. It also requires another supportive tool.

Importantly, for traffic measurement and analysis, the CoralReef and Tcpdump tools cannot be applied as full tools. For traffic analysis and data transferring, other two tools need to support this process. In addition, the L7 filter, Tstat and NetFlow tools are not integrated with windows operating system. L7 filter and Tstat tools do not perform effectively on the traffic measurement and analysis based dynamic environment. But, comparatively, Tstat tool has competitive edge over L7 filter in the processes of traffic measurement. For the purpose of traffic analysis and transferring protocols, the NetFlow is relatively effective; however, it cannot function as effectively in the processes of traffic measurement based dynamic network environments. Furthermore, the Snort tool cannot support the flexibility of traffic measurement and analysis features such as classifying the network usage level, segmenting the data based on its size, the traffic measurement and analyzing the traffic based on various transferring protocols. Critically, Snort does not perform in the low speed networks.

Other tools such as Pandora FSM, Network Monitor and Angry IP scanner are considered as monitoring tools to manage and show the traffic on network ports (Parekh and Patel, 2015). The main advantages of these tools are simple to install and run, uses graphical interfaces and easy to apply. However, there are critical drawbacks of these tools. These tools are not applicable for scalable traffic based features as the tools only monitor and manage available network ports instead of measuring the level of network traffic.

According to (Parekh and Patel, 2015; Gadge and Patil, 2008; Quick and Choo, 2013), Network Monitor is effective for traffic analysis but is not for network traffic measurement as it did not collect the traffic data of the network. Pandora FSM tools however are not applicable on distributed networks as there is a need to provide central server for the specific network. The drawback of Angry IP scanner is that it cannot detect open ports and consider them as filtered whatever the traffic level on these ports.

Findings

Drawing on five key variables, the flexibility of various traffic measurement and analysis tools has been conducted. Those include: (I) compatibility with operating systems, (II) performance of traffic measurement, (III) performance of traffic analysis, (IV) scalable feature ability and (V) transferring protocols flexibility. The following is a discussion of each variable:

- Operating Systems: Tools are compatible with all operating systems that are the Hadoop/MapReduce Snort, Pandora FSM, Network Monitor and Angry IP scanner tools whereas the other tools are suited to UNIX and Linux operating systems rather than Windows
- Performance of traffic measurement: The tools such as Hadoop/MapReduce, Snort, CoralReef, tcpdump, Pandora FSM, Network Monitor, Angry IP scanner and Tstat can manage the traffic measurement processes whereas the NetFlow and L7filter tools cannot be effective for computation of the traffic measurement
- Performance of traffic analysis: Tools such as Hadoop/MapReduce, Snort, NetFlow, L7filter, Pandora FSM, Angry IP scanner and Tstat tools can be applied to handle the traffic analysis processes whereas CoralReef, Network Monitor and tcpdump tools are not used to compute the traffic analysis
- Scalable feature ability: Hadoop/MapReduce method can apply scalable features to measure and analyze the internet traffic measurement and analysis whereas the other tools are fixed i.e., build up tools
- Transferring protocols flexibility: Other tools can transfer the data based on specific transferring protocol whereas Hadoop/MapReduce Pandora FSM, Network Monitor Angry IP scanner and Net Flow tools are compatible to various transferring protocols

The Table 1 illustrates the advantages and disadvantages of various tools of internet traffic measurement and analysis with reference to the tools flexibility as main indicator of the proposed comparison.

Figure 2 shows comparisons between four tools that are Hadoop/MapReduce, Pandora, Network Monitor and Angry IP scanner according to other features scaled from 1-10 for each feature with 10 is the best.

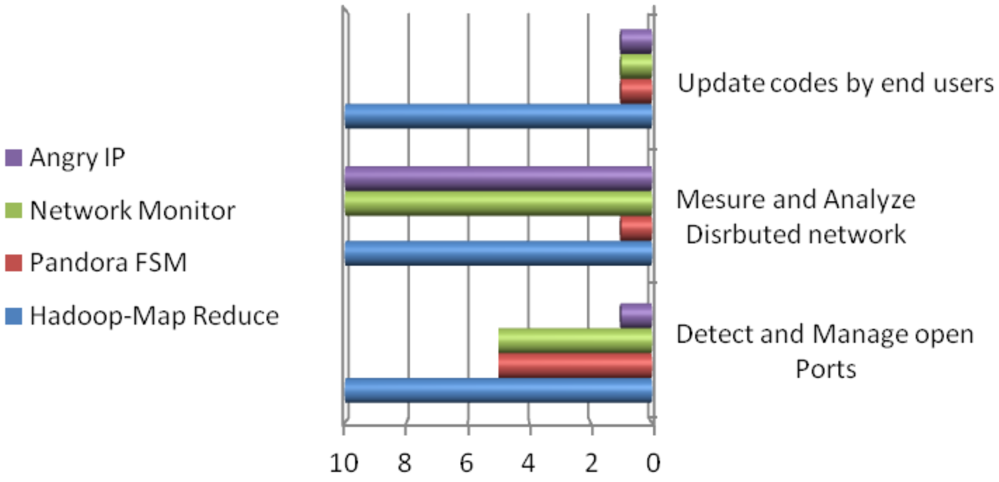

Fig. 2. Comparisons between most effective tools

Table 1. Comparisons of Various approaches of internet traffic measurement and analysis

Tools	Operating system	Traffic measurement	Traffic analysis	Scalable feature ability	Transferring protocols flexibility
Hadoop/Map reduce	All	Use	Use	Use	Use
NetFlow	Not compatible with Windows	Not use	Use	Not use	Use
CoralReef	Not compatible with Windows	Use	Not use	Not use	Not use
Tcpdump	Not compatible with Windows	Use	Not use	Not use	Not use
L7 Filter	Not compatible with Windows	Not use	Use	Not use	Not use
Tstat	Not compatible with Windows	Use	Use	Not use	Not use
Snort	All	Use	Use	Not use	Not use
Pandora FSM	All	Use	Use	Not use	Use
Network monitor	All	Use	Not use	Not use	Use
Angry IP	All	Use	Use	Not use	Use

Recommendations

Hadoop/MapReduce is the most applicable tool for a variety of purposes such as its capacity to be programmed in several programming languages such as C++ and java, its compatibility with numerous operating systems, its capacity to cope with the traffic measurement and analysis so on. In addition, it optimizes the possibility to update the traffic measurement features such as scaling the level of network usage and providing several segmentations of the data. In addition, with the help of Hadoop Map Reduce, the traffic analysis can accomplish dynamically before transferring the data using different transferring protocols such as TCP and UDP. Figure 3 illustrates other advantages of Hadoop/MapReduce to measure and analyzes big data traffic.

Based on main findings of this paper, we would like to recommend the Hadoop/MapReduce tool for the purpose of internet traffic measurement and analysis. By virtue of its flexibility in updating the features of the measurement process, Hadoop/MapReduce can support the scalable measurement and analysis. For example, subject to the data size and network usage, it can support in classifying the network usages and producing effective data segmentation number. In addition to this, while transferring the data through many transferring protocols such as TCP and UDP, the Hadoop/MapReduce provides flexible traffic analysis However, notwithstanding its advantages and strengths, there are many issues Hadoop/MapReduce is yet to handle a number of issues such as classifying the network usage (i.e., low, medium or heavy usage)? Besides, it is yet to determine what the best segmentation number of data is with regard to the networks usage level and data characteristics (i.e., size). Finally, it cannot properly answer what the performance level of transferring time and transferring protocols of the data type are (i.e., text or images)?

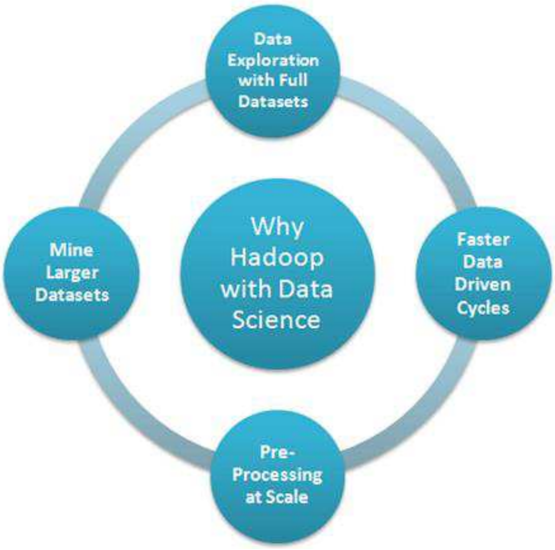
Data Exploration with Full Datasets
Mine Larger Datasets
Why Hadoop with Data Science
Faster Data Driven Cycles
Pre-Processing at Scale

Fig. 3. Advantages of Hadoop/MapReduce

Conclusion

This study investigates various approaches of Internet traffic measurement and analysis. The advantages and disadvantages of the each tool are analyzed based on the implementation flexibility of those tools based with dynamic networks environments. Hadoop/MapReduce approach record competitive advantages over other tools. Hadoop/MapReduce has ability to support the scalable measurement and analysis processes due to its flexibility of update the features of the measurement processes and traffic analysis while transfer the data.

Acknowledgment

The authors would like to acknowledge the assistance provided by the Network and Communication Technology Research Group, FTSM, UKM in providing facilities throughout the research. The research is supported by the Malaysian Government Fundamental Research Grant code FRGS/1/2015/ICT03/UKM/02/2.

Author's Contributions

Lena T. Ibrahim: Conducted critical review, analysis, problem identification and writing of the article.

Rosilah Hassan: Supervisor of the work and reviewed the article.

Kamsuriah Ahmad: Contributed to the reviewing of the article.

Asrul Nizam Asat: Contributed technical, simulation assistance and writing of the article.

Halizah Omar: Contributed in proofreading and overall structure of the article.

Ethics

This article is original and contains unpublished materials. The corresponding author confirms that all of the authors have read and approved the article.

References

Abrahamsson, H., 2008. Internet traffic management. PhD. Thesis, Mälardalen University.

Adami, D., C. Callegari, S. Giordano, M. Pagano and T. Pepe, 2012. Skype-hunter: A real-time system for the detection and classification of Skype traffic. Int. J. Commun. Syst., 25: 386-403. DOI: 10.1002/dac.1247

Ahuja, R.K., T.L. Magnanti and J.B. Orlin, 2015. Network Flows. 1st Edn., BiblioLife, ISBN-10: 297491769, pp: 226.

Alcock, S. and R. Nelson, 2013. Measuring the accuracy of open-source payload-based traffic classifiers using popular Internet applications. Proceedings of the 38th Conference on Local Computer Networks Workshops, Oct. 21-24, IEEE Xplore Press, Sydney, NSW, pp: 956-963. DOI: 10.1109/LCNW.2013.6758538

Angryip, 2014. Angry IP Scanner. http://angryip.org

Arlitt, M. and C. Williamson, 2005. An analysis of TCP reset behaviour on the internet. ACM SIGCOMM Comput. Commun. Rev., 35: 37-44. DOI: 10.1145/1052812.1052823

Awduche, D., A. Chiu, A. Elwalid, I. Widjaja and X. Xiao, 2002. Overview and principles of Internet traffic engineering. Internet RFC.

Awduche, D., J. Malcolm, J. Agogbua, M. O'Dell and J. McManus, 1999. Requirements for traffic engineering over MPLS. Internet RFC.

Bhattacharyya, S., C. Diot, J. Jetcheva and N. Taft, 2001. Pop-level and access-link-level traffic dynamics in a tier-1 POP. Proceedings of the ACM SIGCOMM Internet Measurement Workshop, Nov. 01-02, Burlingame, CA, USA, pp: 39-53. DOI: 10.1145/505202.505209

Brownlee, N. and K.C. Claffy, 2002. Understanding Internet traffic streams: Dragonflies and tortoises. IEEE Commun. Magazine, 40: 110-17. DOI: 10.1109/MCOM.2002.1039865

Brownlee, N. and K.C. Claffy, 2004. Internet measurement. IEEE Internet Comput., 8: 30-33.

Brownlee, N., 1997. Traffic flow measurement: Experiences with NeTraMet. The University of Auckland.

Carela-Espanol, V., P. Barlet-Ros, A. Cabellos-Aparicio and J. Sole-Pareta, 2011. Analysis of the impact of sampling on NetFlow traffic classification. Comput. Netw., 55: 1083-1099. DOI: 10.1016/j.comnet.2010.11.002

Dittrich, J. and J.A. Quiané-Ruiz, 2012. Efficient big data processing in Hadoop MapReduce. Proc. VLDB Endow., 5: 2014-2015. DOI: 10.14778/2367502.2367562

Dong, S., D. Zhou, W. Zhou, W. Ding and J. Gong, 2013. Research on network traffic identification based on improved BP neural network. Applied Math. Inform. Sci., 7: 389-398. DOI: 10.12785/amis/070148

Finamore, A., M. Mellia, M. Meo, M. Munafo and D. Rossi, 2011. Experiences of internet traffic monitoring with tstat. IEEE Netw., 25: 8-14. DOI: 10.1109/MNET.2011.5772055

Finamore, A., M. Mellia, M. Meo, M.M. Munafo and D. Rossi, 2010. Live traffic monitoring with tstat: Capabilities and experiences. Proceedings of the 8th International Conference on Wired/Wireless Internet Communication, Jun. 1-3, Luleå, Sweden, pp: 290-301. DOI: 10.1007/978-3-642-13315-2_24

Fortz, B. and M. Thorup, 2000. Internet traffic engineering by optimizing OSPF weights. Proceedings 19th Annual Joint Conference of the IEEE Computer and Communications Societies, Mar. 26-30, IEEE Xplore Press, Tel Aviv, pp: 519-528. DOI: 10.1109/INFCOM.2000.832225

Fortz, B. and M. Thorup, 2002. Optimizing OSPF/IS-IS weights in a changing world. IEEE J. Selected Areas Commun., 20: 756-767. DOI: 10.1109/JSAC.2002.1003042

Fusco, F. and L. Deri, 2010. High speed network traffic analysis with commodity multi-core systems. Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement, Nov. 01-03, Melbourne, Australia, pp: 218-224. DOI: 10.1145/1879141.1879169

Gadge, J. and A.A. Patil, 2008. Port scan detection. Proceedings of the 16th IEEE International Conference on Networks, Dec. 12-14, IEEE Xplore Press, New Delhi, pp: 1-6. DOI: 10.1109/ICON.2008.4772622

Gebert, S., R. Pries, D. Schlosser and K. Heck, 2012. Internet access traffic measurement and analysis. Proceedings of the 4th International Conference on Traffic Monitoring and Analysis, (TMA' 12), Vienna, Austria, pp: 29-42. DOI: 10.1007/978-3-642-28534-9_3

Grajzer, M., M. Koziuk, P. Szczechowiak and A. Pescape, 2012. A multi-classification approach for the detection and identification of eHealth applications. Proceedings of the 21st International Conference on Computer Communications and Networks, Jul. 30-Aug. 2, IEEE Xplore Press, Munich, pp: 1-6. DOI: 10.1109/ICCCN.2012.6289268

Grimaudo, L., M. Mellia and E. Baralis, 2012. Hierarchical learning for fine grained internet traffic classification. Proceedings of the 8th International Wireless Communications and Mobile Computing Conference, Aug. 27-31, IEEE Xplore Press, Limassol, pp: 463-468. DOI: 10.1109/IWCMC.2012.6314248

Gunnar, A., H. Abrahamsson and M. Soderqvist, 2005. Performance of Traffic engineering in operational IP networks: An experimental study. Proceedings of the 5th IEEE International Workshop on IP Operations and Management, Oct. 26-28, Barcelona, Spain, pp: 202-211. DOI: 10.1007/11567486_21

Hasib, M. and J.A. Schormans, 2003. Limitations of passive and active measurement tools in packet networks.

John, W., 2008. On measurement and analysis of internet backbone traffic. Chalmers University of Technology.

Kadam, Y.V. and V. Dhore, 2010. A study on scalable internet traffic measurement and analysis with hadoop. Int. J. Eng. Comput. Sci., 2: 3187-3190.

Karagiannis, T., A. Broido, N. Brownlee, K. Claffy and M. Faloutsos, 2004. Is P2P dying or just hiding?. Proceedings of the Global Telecommunications Conference, (GTC' 04), IEEE Xplore Press, pp: 1532-1538. DOI: 10.1109/GLOCOM.2004.1378239

Karagiannis, T., K. Papagiannaki and M. Faloutsos, 2005. BLINC: Multilevel traffic classification in the dark. Proceedings of the Conference on Applications, Technologies, Architectures and Protocols for Computer Communications, Aug. 22-26, Philadelphia, PA, USA, pp: 229-240. DOI: 10.1145/1080091.1080119

Keys, K., D. Moore, R. Koga, E. Lagache and M. Tesch, 2001. The architecture of CoralReef: An internet traffic monitoring software suite. Center for Applied Internet Data Analysis.

Lee, Y. and Y. Lee, 2013. Toward scalable internet traffic measurement and analysis with hadoop. ACM SIGCOMM Comput. Commun. Rev., 43: 5-13. DOI: 10.1145/2427036.2427038

Lee, Y., W. Kang and Y. Lee, 2011. A hadoop-based packet trace processing tool. Proceedings of the 3rd International Workshop on Traffic Monitoring and Analysis, (TMA' 11), Vienna, Austria, pp: 51-63. DOI: 10.1007/978-3-642-20305-3_5

Liu, J., F. Liu and N. Ansari, 2014. Monitoring and analyzing big traffic data of a large-scale cellular network with Hadoop. IEEE Netw., 28: 32-39. DOI: 10.1109/MNET.2014.6863129

Mehra, P., 2012. A brief study and comparison of snort and bro open source network intrusion detection systems. Int. J. Adv. Res. Comput. Commun. Eng., 1: 383-386.

Microsoft.com, 2012. https://blogs.technet.microsoft.com/messageanalyzer/2012/09/17/meet-the-successor-to-microsoft-network-monitor/

Parekh, N.B. and A.J. Patel, 2015. A survey on internet traffic measurement and analysis. Comput. Eng. Intelli. Syst., 6: 75-80.

Poppe, F., S. Van den Bosch, P. de La Vallée-Poussin, H. Van Hove and H. De Neve *et al*., 2000. Choosing the objectives for traffic engineering in IP backbone networks based on quality-of-service requirements. Proceedings of the 1st COST 263 International Workshop on Quality of Future Internet Services, Sept. 25-26, Berlin, Germany, pp: 129-140. DOI: 10.1007/3-540-39939-9_11

Pries, R., F. Wamser, D. Staehle, K. Heck and P. Tran-Gia, 2009. Traffic measurement and analysis of a broadband wireless internet access. Proceedings of the IEEE 69th Vehicular Technology Conference, Apr. 26-29, IEEE Xplore Press, Barcelona, pp: 1-5. DOI: 10.1109/VETECS.2009.5073890

Quick, D. and K.K.R. Choo, 2013. Digital droplets: Microsoft SkyDrive forensic data remnants. Future Generat. Comput. Syst., 29: 1378-1394. DOI: 10.1016/j.future.2013.02.001

Rafeeq, U.R., 2003. Intrusion Detection Systems with Snort: Advanced IDS Techniques Using Snort, Apache, MySQL, PHP and ACID, Prentice Hall Professional, Upper Saddle River, NJ, ISBN-10: 0131407333, pp: 263.

Ramakrishnan, K.G. and M.A. Rodrigues, 2001. Optimal routing in shortest-path data networks. Bell Labs Tech. J., 6: 117-138. DOI: 10.1002/bltj.2267

Roesch, M., 1999. Snort-lightweight intrusion detection for networks. Proceedings of the 13th USENIX Conference on System Administration, Nov. 7-12, Seattle, Washington, USA, pp 229-238.

So-In, C., 2009. A survey of network traffic monitoring and analysis tools. Washington University.

So-In, C., 2014. A survey of network traffic monitoring and analysis tools. Washington University, St. Louis.

Sridharan, A., R. Guerin and C. Diot, 2003. Achieving near-optimal traffic engineering solutions for current OSPF/IS-IS networks. IEEE/ACM Trans. Netw., 13: 234-247. DOI: 10.1109/TNET.2005.845549 .

Yuan, D., 2001. Optimization models and tools for communication network design and routing. PhD Thesis, Linkpings Universitet.