Comparative Analysis of Internet Traffic Identification Methods

Kokila S
Assistant Professor, Department of Computer Science, Maharaja Arts & Science College, Coimbatore-6.
Email: kokisakthi4444@gmail.com
Sathish A
Assistant Professor, Department of Computer Science, Maharaja Arts & Science College, Coimbatore-6.
Email: asathish67@gmail.com
Shankar R
Assistant Professor, Department of Computer Science, Chikkanna Government Arts College, Tiruppur.
Email: shankarchac@gmail.com

----------ABSTRACT----------

The area of Internet traffic measurement has advanced enormously over the last couple of years. This was mostly due to the increase in network access speeds, due to the appearance of bandwidth-hungry applications, due to the ISPs' increased interest in precise user traffic profile information and also a response to the enormous growth in the number of connected users. These changes greatly affected the work of Internet Service Providers and network administrators, which have to deal with increasing resource demands and abrupt traffic changes brought by new applications. This survey explains the main techniques and problems known in the field of IP traffic analysis and focuses on application detection. This paper addresses the network traffic aspects of Internet to reduce traffic

Keywords**: Measurement, Application Identification, Traffic Analysis, Classification**.

I. INTRODUCTION

Network traffic measurement has recently gained more interest as an important network-engineering tool for networks of multiple sizes. The traffic mix flowing through most long-haul links and backbones needs to be characterized in order to achieve a thorough understanding of its actual composition. Different applications (traditional ones such as Web, malicious others such as worms and viruses or simply hype such as P2P) affect the underlying network infrastructure. New business and settlement models may be reached between content and trans-port providers once a clear traffic understanding is achieved.

A.

In broader terms, measurement strategies can be seen as an essential tool for identifying anomalous behavior, for the design and validation of new traffic models, for offering highly demanded services, as well as for helping seasonal activities such as upgrading network capacity or eventually for usage-based pricing.

But first, it is very important to differentiate between network measurement and application identification: the former is about data gathering and counting. Traffic identification, however, is inherent to traffic classification, since one may not classify before identification. According to, traffic measurements can be divided in active and passive measurements; and can also be divided in online and offline strategies. In the case of online measurement, the analysis is performed while the data is captured; while in offline measurements, a data trace is stored and analyzed later.

1.1 Active versus Passive Measurements

Active measurement is defined as measurement obtained through injected traffic. In the case of active monitoring several probe packets are sent continuously across the network to infer its properties. Active measurements are mainly used for fault and vulnerability detection and network or application performance tests.

However, it may not be always suitable to reveal network characteristics as influenced by users, due to the fact that active measurement sends packets independently of user behavior and therefore changes the network metrics it is trying to measure in the first place. Passive measurement is defined as measurement of existing traffic without injecting traffic. Passive techniques are carried out by observing network packets and connections. A flow is defined as a set of packets that share origin and destination addresses, origin and destination ports, transport protocol and are observed within a time-frame (this is configurable).

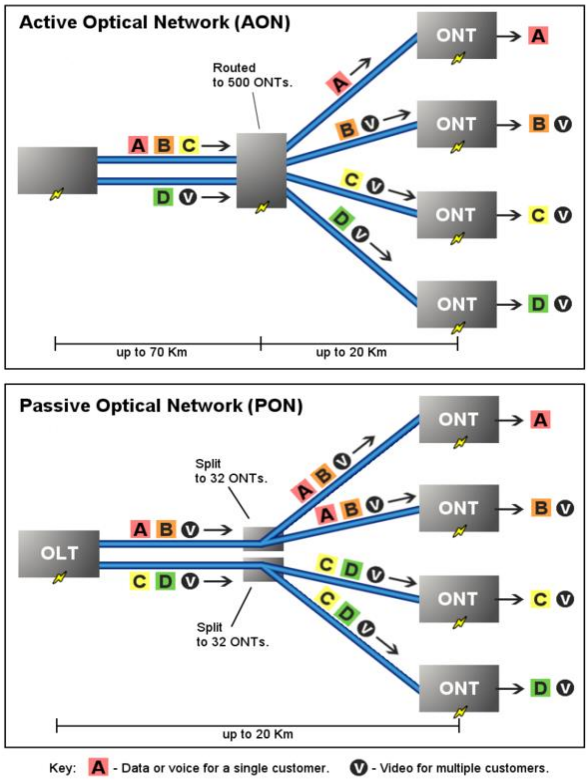

Figure 1 Active & Passive Traffic Identification

1.2 Packet-Level Passive Measurements

At a microscopic level, measurements are performed on each packet traveling across the measurement point. The information collected can be very fine-grained. Examples of relevant collected information are source and destination IP address, source and destination port numbers, packet sizes, protocol numbers and specific application data. There are several packet capture tools (sniffers) freely available, most of which rely on the libpcap library.

TCP dump is a commonly used tool that allows one to look closer at network packets and make some statistical analysis out of the trace files. Ethereal or Wireshark, as currently known, adds a user-friendly GUI to TCPdump and includes many traffic signatures that can be used for accurate, payload-based application identification. SNORT is a tool for real-time traffic analysis and packet logging, capable of performing content searching/matching and detecting many types of network security attacks.

A set of other packet-related tools may be found in the Internet. There are three possible hardware combinations for packet capture.

- Cable splitter
- Port mirroring
- Firewall or Traffic shaper.

This may increase packet delay, but the equipment will not change the packet's contents. Depending on which part of the captured data will be stored, some processing may be required in the packet capture machine such as converting some fields or hiding certain information for privacy concerns. Next, the data may be stored in a local or remote database for scalability and proper data management and will be available for traffic management analysis requests.

1.3 Flow-Level Passive Measurements

At a macroscopic level, measurements are performed on flow basis. In this case, aggregation rules are necessary to match packets into flows. Collected data include the number of flows per unit of time, flow bit rate, flow size and flow duration. Examples of commonly used tools that deal with flows are Cisco's NetFlow (the de facto standard) and Juniper's JFlow.

Cisco was the first to come up with and implement a flow-level capture solution. NetFlow provides a set of services for IP applications, including network traffic accounting, usage-based network billing, network planning, security control, Denial of Service (DoS) monitoring capabilities, and network monitoring. It is currently seen as the most important technology for measuring and exporting traffic flows.

Although NetFlow v5 provides many fields of information, in practice many programs fail to correctly fill all its fields. Consequently, the only systematically utilized (i.e., correctly fulfilled) and therefore dependable fields are: Source IP, Destination IP, Source Port, Destination Port, Layer 4 Protocol, Packet Count, Byte Count, Start Time and End Time.

JFlow also provides a similar set of functionalities and supports NetFlow's export formats. Actually, most software developers of flow collectors along with the leading companies in the router-related industry are working jointly within the IETF to build the records representation known as IPFIX.

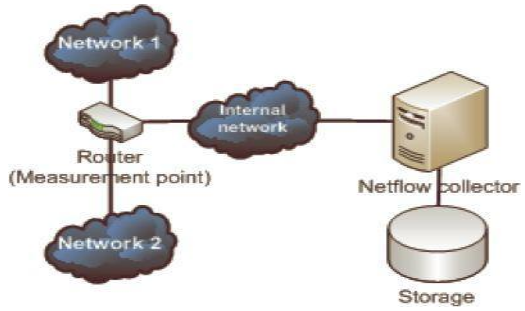

Figure 2 Flow Measurement Topology

II. STATE-OF-THE-ART IN FLOW ANALYSIS

One of the main problems with passive measurement is dealing with a massive amount of data, since the volume of captured data can become very large on high-capacity links. Additionally, the network manager should make important design decisions on how to cope with different granularity levels in order to gather useful information for network traffic engineering.

Essentially, there is a broad avenue for future research in this field, spanning all the way from defining strategies – to sampling - for dealing with the huge amount of network traffic data.

The authors argue that the system can accept measurement tasks, locate required measurement facilities and fulfill these tasks. All experiments and simulation results point out that p2p-based measurement system are very promising, but no results of a practical measurement were shown.

As an important step for a cautious deployment of any new technique for network management, the research work in analyses the loss of information caused by TCP (Transmission Control Protocol) traffic aggregation in flows. A tool called FLOW-REDUCE was implemented to reconstruct TCP connection summaries from NetFlow export data (named flow connections). The problem is that NetFlow breaks flow information in 5-minute flows, thus possibly breaking a given captured flow into many flows. To study how accurately information is produced by FLOW-REDUCE, TCP connection summaries are reconstructed from packet traces using the BRO tool.

2.1 Classification: Volume, Duration, Rate and Burstiness

In a similar approach to, the research work in per-forms a multi-scale and multi-protocol analysis to explore the persistency (volume, duration, rate and burstiness) properties of those flows that contribute the most to bandwidth utilization (the flows are called elephants or heavy hitters). The authors argue that knowing the persistency features of heavy hitters and understanding their underlying causes is crucial when developing traffic-engineering tools that focus primarily on optimizing system performance for elephant flows.

The main difficulty that arises when studying the persistency properties of flows is that the available measurements are either too fine-grained to perform large-scale studies (i.e., packet-level traces) or too coarse-grained to extract the detailed infor-mation necessary for the particular purpose (i.e., NetFlow traces or MIB/SNMP data).

2.2 Traffic Characterization

Many network traffic modeling research papers initiate with a traffic analysis approach before proposing any analytical one. Therefore, one can take advantage of this procedure to gain some important knowledge on the most common types of analysis for network traffic.

In, the authors developed a new framework for analyzing and modeling network traffic that reaches beyond aggregation by incorporating connection-level information. A careful study of many traffic traces acquired in different networking situations reveals that traffic bursts typically arise from just a few high-volume connections that dominate all others.

Queuing experiments suggest that the alpha component dictates the tail queue behavior for large queue sizes, whereas the beta component controls the tail queue behavior for small queue sizes. The potential causes of burstiness might range from the transient response to re-routing, the transient response to start/stop of connections, the TCP slow-start peculiarities to the heterogeneity in bottleneck links for passing flows.

2.3 User Behavior

Due to the increased complexity and processing power required for performing user behavior analysis from packet traces and the restrictions on payload data usage in force in many countries, recent studies focus on the analysis of flow traces or connection-level behavior. Effective traffic analysis will provide statistically sound general network profiles and application-specific behavior.

Volume analysis has already been considered a very insensitive method for anomaly detection, although it may reveal few anomalies faster and easier. This is due to the aggregative and consequently information destructive characteristic of volume statistics.. Flow-level analysis saves on processing resources and has also shown to be useful for anomaly detection. Flow and volume analysis together should form a good methodology for anomaly detection, considering that both detect rather disjoint sets of

anomalies. Furthermore, some previous work has been done on IP traffic characterization and focused on understanding statistical properties of packet and flows at the network and transport layers.

In this area some works are considered understanding seasonal traffic volumes, user connection durations, traffic growing trends, packet arrival processes, self similar (fractal) behavior and traffic matrix estimation. This crucial information has been used both by ISPs for network dimensioning and resource provisioning and by the Internet research community for an in-depth understanding of the current Internet traffic state and protocol design.

.

III. TRAFFIC CLASSIFICATION RESULTS

When studying a traffic classification technique with real traces, it is important to have a baseline for traffic classification that will be used as a trustable reference. This can be achieved by manual classification of traffic traces, by the use of active measurement or by another method (e.g., a payload classification tool) that was proved to have a high accuracy. It is not yet clear how recent this proof should be, since new applications keep emerging daily.

Due to the problems discussed above, validation remains a difficult task. Comparing the accuracy and completeness of the algorithms based on different measurements with potentially different reference classifications, over different networks is not straightforward. In addition, some papers only evaluated the accuracy and not the completeness of their methods. In addition, each paper uses a different traffic source for evaluation, which makes a trustable comparison unreliable. The next section presents a comparison of different traffic identification methods based on the results of their authors.

A. Comparison of traffic identification methods

The authors develop their own byte signatures to establish a baseline and use it for validating the proposed connection pattern based classification method. As shown in Table I, the Blinc algorithm is able to recognize the main application types. For their trace, the algorithm performs better in terms of accuracy than completeness (conservative detection).

For comparison, the next method we study is the Bayesian Analysis. First, this method needs to be trained with a data set that was previously classified e.g. manually. Then, the method is tested on a different data set. The authors investigate the accuracy of the approach but do not address completeness, which is very important to validate the relevance of the accuracy metric, as explained before. In this method, also shown in Table I, the accuracy of the P2P file sharing traffic is much lower than of other applications. This is in line with the fact that P2P applications are diverse and their main characteristics are difficult to grab, especially with the packet metrics utilized by the Bayesian method.

The method proposed by, the “On the Fly” algorithm, uses flow clustering and also uses learning for cluster labeling, but it reads only the first few packet headers in each connection. The accuracy of the classification methods is also summed up in Table I. Here, the authors use payload analysis as a reference or baseline. According to the results, the On the Fly method works roughly as accurately as the Bayesian method even though it relies on significantly less and simpler input.

Table I shows that the algorithms performed roughly well on the analyzed traces. On the other hand, the fact that all three methods use heuristics implies that some fine tuning work may be needed to fit the methods to other traces or new applications.

For comparison purposes, the results from a byte signature based analysis (referred to as DPI) are also shown in Table I. This analysis was made possible by the manual identification of flows for comparison. When comparing the payload analysis with other identification techniques, it becomes clear that these techniques achieve slightly better results than those shown by the DPI.

First, these techniques can capture the behavior of an application, sometimes finding new applications that still had no payload signature but behaved similarly to applications of the same type. Second, this comparison is unfair, since most papers with behavioral analysis only considered TCP.

This is because, for an independent identification technique (not combined with other technique), a high accuracy with a low completeness is just as good as a low accuracy with a good completeness.

Since the completeness is unknown, the meaningfulness of their accuracy also becomes unknown. One way of getting around the uncertainty provided by the heuristics is to run more algorithms in parallel, compare their results, and conclude the final application classification decision based on the result of the comparison, as introduced by (with their own algorithms). This approach also has the advantage that mismatching classification results are recognized automatically. Furthermore, such traffic may be dumped separately for further analysis and the knowledge gained can be incorporated into the algorithms.The accuracy and the completeness of the classification methods on different application types compared to their combined classification method can be seen in table

The joint application – and the comparison run on the same input data that some methods are stronger in accuracy and others provide more complete results (see Table I). As a consequence, the application classification decision is a trade-off between the amount of traffic left unknown and the bigger likelihood of erroneous classification.

Application	Metric	BLINC	Bayesian	On the Fly	Payload Analysis
WWW	Completeness	69-97%	-	-	134%
	Accuracy	98-100%	99.27	-	91%
.HTTP	Accuracy	-	-	99%	-
.HTTPS	Accuracy	-	-	81.8%	-
Mail	Completeness	78-95%	-	-	78%
	Accuracy	85-99%	94.78%	-	97%
.SMTP	Accuracy	-	-	84.5%	-
.pom3	Accuracy	-	-	0%	-
.pop3S	Accuracy	-	-	89.8%	-
Bulk file transfer(FTP)	Completeness	95%	-	-	26%
	Accuracy	98%	82.25%	87%	99%
NNTP	Accuracy	-	-	99.6%	-
Chat	Completeness	68%	-	-	76%
	Accuracy	98%	-	-	97%
Network Management	Completeness	85-97%	-	-	75%
system	Accuracy	88-100%	-	-	95%
Services(server)	Accuracy	-	63.68%	-	-
Database	Accuracy	-	86.91%	-	-
Multimedia/streaming	Completeness	-	-	-	3%
	Accuracy	-	80.75%	-	98%
Peer-to-peer	Completeness	84-90%	-	-	61%
(File sharing)	Accuracy	96-98%	36.45%	-	99%
.Edonkey	Accuracy	-	-	84.2%	-
SSH	Accuracy	-	-	96.92%	-
Tunneled	Completeness	-	-	-	120%
	Accuracy	-	-	-	10%

Table 1: Accuracy and Completeness of the Identification Method

Many techniques for network management and application identification do exist, but some suffer from legal problems (signature-based packet payload analysis) while others (inference-based) only identify a few applications correctly. Well-known-ports are no longer an answer, since many applications, especially those with a high network volume (e.g., P2P file sharing), bypass the rules and use known ports of other services.

Payload-based schemes are very time-consuming, therefore should not be utilized in real-time in high-speed links, except when using high cost specialized hardware in specific network links (up to 1Gbps).

Flow-based schemes (inference) lose information, and even these require sampling on very high speed links, depending on the routers. Some authors claim their inference-based methods achieve high efficiency and precision, but it greatly varies with the traffic pattern studied.

IV. PROBLEMS AND DIRECTIONS

Based on this survey of traffic identification papers, the authors identified some issues that still remain open, the best level of detail for measurements is still not defined. This leads to a multidimensional problem constrained by existing equipment for measurements and main traffic characteristics. From the research point of view, the problem is to find the minimal amount of data that needs to be measured in order to classify applications. However, storing the minimal amount of data may not be the best solution, since additional data may be needed to validate results. Furthermore, in practice, measured results usually raise additional questions and existing extra measurement may help in prompt replies.

One way to deal with this problem is to apply sampling or other filtering techniques, e.g. measure the traffic of selected subscribers. It is not clear how much sampling can be used to keep a certain level of accuracy. It is also not clear how much information is lost given a certain sampling approach.

V. CONCLUSIONS

Internet measurement is a very dynamic and wide field; all the time new approaches to network management, application profiling and traffic modeling are proposed, each analyzing a different aspect. Packet-based application inference has some issues that may not be circumvented technologically. Flow-based application inference is still an incipient field of study, despite the many papers on the subject. Using present day research, none of them achieve a high accuracy with a high precision in a broad range of applications. Further study is required on a new technique for dependable application detection.

VI. REFERENCES

[1] Azzouna, Nadia Ben and Guillemin, Fabrice, *Analysis of ADSL Traffic on an IP Backbone Link*, IEEE IEEE Global Telecommunications Conference 2003, San Francisco, USA, December 2003.

[2] Sullivan, Mark, *Surveys: Internet Traffic Touched by Youtube*, Light Reading, http://www.lightreading.com/document.asp?doc_id= 115816, January 2007.

[3] Cho, Kenjiro; Fukuda, Kenshue; Esaki, Hiroshi and Kato, Akira, *The Impact and Implications of the Growth in Residential User-to-User Traffic*, ACM SIGCOMM 2006, Pisa, Italy, September 2006.

[4] Balachandran, Anand; Voelker, Geoffrey M.; Bahl, Paramvir and Ragan, P. Venkat, *Characterizing user behavior and network performance in a public wireless LAN*, Proceedings of the 2002 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, pp. 195-205, 2002.

[5] Cherry, S., *The VoIP Backlash*, IEEE Spectr., October 2005, http://spectrum.ieee.org/oct05/1846.

[6] Cisco IOS NetFlow, *Introduction to Cisco IOS NetFlow - A Technical Overview*, White Paper, Last updated: October 2007,