Broadband network traffic characterization and classification using a multivariate statistical method

Raimir Holanda
Professor University of Fortaleza raimir@unifor.br

José Everardo Bessa Maia
Professor University of Fortaleza jmaia@unifor.br

Gabriel Paulino
Master student at University of Fortaleza.
gabrieljr@yahoo.com

Resumo

O comportamento do tráfego em backbones de redes está constantemente sendo alterado devido a altas demandas requeridas por um determinado serviço, ataques à rede, surgimento de novos serviços, entre outros. Embora a caracterização e classificação de tráfego de rede seja uma tarefa bem conhecida, ela deve ser necessariamente efetiva em situações anômalas de tempo real, com o propósito de ajudar a manter a rede com bom desempenho. Abordagens clássicas como o uso de mecanismos de inteligência artificial, foram modificadas a fim de tentar alcançar essas exigências. Entretanto, tais ajustes os deixam lentos, necessitando de muitos recursos e da participação constante do administrador da rede. Este trabalho apresenta uma investigação de uma metodologia para caracterizar e classificar padrões em um tráfego de rede em banda larga. Esta metodologia é baseada na análise de agrupamento, um método estatístico multivariado empregado para descobrir associações e estruturas em dados. A análise de agrupamento possibilita a extração de padrões dos fluxos de dados.

Palavras-chave: *Caracterização e classificação de tráfego. Estatística multivariada.*

Abstract

Network traffic behavior is constantly changing due to issues as high service demand on a given service, network attacks, emergence of new services, among others. Although network traffic characterization and classification is a well-known task, it must mainly be effective in real-time anomalous situations in order to help to keep the network with good performance. Classical approaches such as the use of artificial intelligence mechanisms have been modified in order to attempt requirements. However, in general, such adaptations are slow, need of many resources and a constant participation of the network administrator. This work presents an investigation of a methodology for characterize and classify patterns into broadband network traffic. This methodology is based on flow clustering analysis, a multivariate statistical method employed for discovering associations and structures in collected data. The clustering analysis enables the extraction of patterns of data flows.

Keywords: *Traffic characterization and classification. Multivariate statistic.*

1 Introduction

Network traffic measurement enables the monitoring of its behavior and also allows the diagnosis of its current state. The behavior of network-wide traffic is constantly changing due to issues as high demand on a given service, network attacks, emergence of new services, among others. Identifying and understanding network traffic is an essential task for keeping the good performance and predicting bandwidth requirements.

In the last years, a new tendency for traffic modeling has emerged based on multivariate statistics such as Factor Analysis (LEDYARD and LEWIS,1973) Principal Component Analysis (LAKHINA et al., 2004) and Cluster Analysis (TAYLOR and ALVES-FOSS, 2002). Multivariate statistics are appropriate for any data set where multiple measurements are taken with possible correlations between the measurements. Multivariate techniques in general, account for the correlation structure of the variables being analyzed often yielding a more complete picture of the analysis results than if the variables had been analyzed separately (JOHNSON, 1998).

The Cluster Analysis technique consists of identifying data flow clusters based on a collection of variables previously established. By means of those clusters becomes possible the extraction of patterns of flows that should represent normal behaviors. Network-wide traffic in general is composed by flows that frequently change along the time. In this way, clustering analysis sounds like an attractive technique that can be applied to model a number of traffic flow on backbone networks.

In this paper we propose, for Internet traffic, a novel semantic characterization and a new approach to classify this traffic based on flow clustering spectrum. We demonstrated that behind the great number of flows in a high-speed link, there is not so much variety among them and clearly they can be grouped into a set of few clusters.

2 Related work

There are many published papers dealing with packet traffic characterization. However, most of them are focused on traffic characteristics such as inter packet time, traffic intensity, packet length, etc; and they do not take into consideration the semantic of flows. We understand by semantic characterization the analysis of traffic characteristics including the inter packet time and some of the most important fields (source and destination address, port numbers, packet length, TCP flags, etc) of the TCP/IP headers content.

In the last years, a number of published works have studied some traffic characteristics such as flow size, flow lifetime, IP address locality, and IP address structure. For instance, in (GUO amd MATTA, 2001) flow size distribution is studied, introducing a flow classification based on number of bytes, i.e, *mice* or *elephants*. In (BROWNLEE and CLAFFY, 2002) flows are classified by lifetime, demonstrating that most flows are very short.

For temporal and spatial locality of reference in streams of requests arriving at Web servers, Almeida, Bestavros, Crovella, and Oliveira, have proposed models that capture both properties (ALMEIDA et al., 1996). Kohler, Li, Paxson, and Shenker (KOHLER et al., 2002), have investigated the structure of addresses contained in IP traffic. All these studies show important characteristics of the traffic, but in our opinion, more semantic aspects of flows are required for a useful traffic characterization.

Clustering methods are described and applied in (MARCHETTE, 1999) to network data. These methods were used to determine activity patterns and whether current activity matches these patterns in order to determine when there is abnormal activity on the network.

An structural analysis of network traffic flows is proposed by Lakhina et al. (2004). As network traffic arises from the superposition of Origin-Destination (OD) flows, they argue that the understanding of OD flows becomes essential for addressing a wide variety of problems, including traffic engineering, traffic matrix estimation, capacity planning, forecasting and anomaly detection.

Lakhina et al. propose the use of sampled flow measurements in an IP network for detecting and understanding network-wide traffic anomalies. This work characterizes a range of network-wide by aggregating sampled flow achieved at the origin-destination (OD) level. In contrast to, the use of cluster analysis in our work takes account not only OD flows, but also extra characteristics found in those flows.

A flow characterization approach that incorporates semantic characteristics of flows was proposed by Holanda (HOLANDA, 2005). Using clustering techniques, he demonstrated that behind the great number of flows in a high speed link, there is not so much variety among then. In addition, traces captured from different links showed similar behavior. Based on the evidence that many flows can be grouped into few clusters, a pattern was extracted. Such pattern means a dataset storing the most common classes of flows. That approach was applied to develop a method for packet trace compression in (HOLANDA et al., 2005).

Such method achieved compression ratio around 3%, reducing the file size, for instance, from 100 MB to 3 MB. Although the proposed method defines a lossy compressed data format, it preserves important statistical properties present into original trace.

3 Cluster analysis technique

The cluster analysis technique encompasses a number of different algorithms and methods for grouping objects of similar kind into respective categories (KAUFMAN and ROUSSEEUW, 1990). In this work, the objects are TCP/IP flows

and are defined as a set of packets which share the following fields: source IP address, destination IP address, source port, destination port and protocol.

A general question facing researchers in many areas of inquiry is how to organize observed data into meaningful structures. In other words, cluster analysis is an exploratory data analysis tool which aims at partitioning the components into groups so the members of a group are as similar as possible and different groups are as dissimilar as possible (JAIN, 1991).

Statistically, this implies that the intragroup variance should be as small as possible and intergroup variance should be as large as possible. Each cluster thus describes, in terms of data collected, the class to which its members belong; and this description may be abstracted through use from the particular to the general class or type.

Thus, cluster analysis is a tool of discovery. It may reveal associations and structure in data which, though not previously evident, nevertheless are sensible and useful once found. The results of cluster analysis may contribute to the definition of a formal classification scheme.

A number of clustering techniques have been described in the literature. These techniques fall into two classes: *hierarchical* and *nonhierarchical*. In nonhierarchical approaches, one starts with an arbitrary set of k clusters, and the members of the clusters are moved until the intragroup variance is minimum. There are two kinds of hierarchical approaches: *agglomerative* and *divisive*. In the agglomerative hierarchical approach, given n components, one starts with n clusters (each cluster having one component). Then neighboring clusters are merged successively until the desired number of clusters is obtained. In the divisive hierarchical approach, on the other hand, one starts with one cluster (of n components) and then divides the cluster successively into two, three, and so on, until the desired number of clusters is obtained.

4 Flow characterization

Network traffic measurements provide a mean to understand the behavior of different kinds of networks. Using specialized network measurement hardware or software, a network researcher can collect detailed information about the transmission of packets on the network,including their time structure and content. The measurements used here are related with TCP/IP traffic into backbone networks. With detailed packet-level measurements, and some knowledge of the Internet Protocol stack, for example, it is possible to obtain significant information about the structure of an Internet application or the behavior of an Internet user.

4.1 Traces investigated

The Off-line analysis carried along this work are based on traces captured on passive mode from many sites in 2005. One of them was an OC-3 (155 Mbps) link trace. This link connects the Scientific Ring of Catalonia to RedIRIS (Spanish National Research Network) (RedIRIS, 2005) as shown in Figure 1. The trace was collected by a hardware-based measurement tool on passive mode. This not sanitized trace is a collection of packets flowing in one direction of the link containing a timestamp and the first 40 bytes of the packet. For our analysis, we have used only the output link. The RedIRIS is a high performance communication network created in 1993 and nowadays it joins more than forty research institutions.

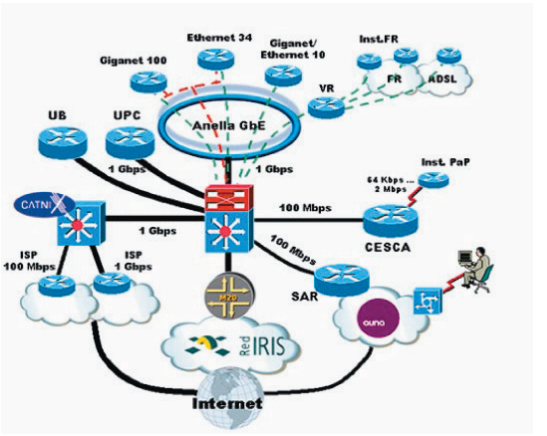

Figure 1: RedIRIS topology

Furthermore, in our study we employ publicly available archive of traces collected and maintained by the National Laboratory for Applied Network Research (NLANR, 2005). We downloaded traces collected from the following sites: Colorado State University (COS), Front Range GigaPOP (FRG), University of Buffalo (BUF), and Columbia University (BWY).

In all cases, the traces were stored using the TSH packet header format. For *.tsh* files the header size is 44 bytes: 8 bytes of timestamp and interface identifier, 20 bytes of IP, and 16 of TCP, as shown in Figure 2. No IP or TCP options are included and the packet payload is also not stored. Apart of using a 5-tuple of fields which define a flow, we have used the following fields into our clustering analysis: TCP Flags, total length, and timestamp (see dark fields).

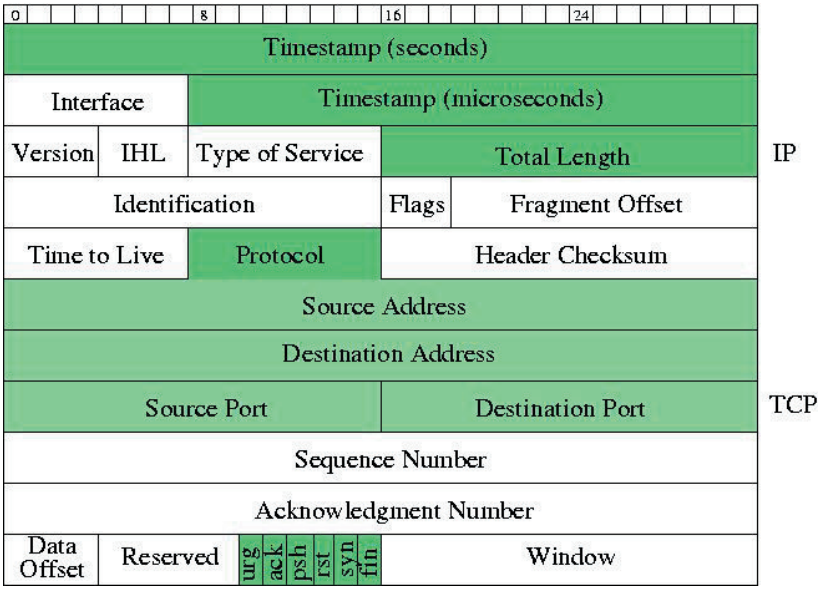

Figure 2: TSH header data format

4.2 Specification of flow clustering

The specification presented here models the traffic at flow level. As we have seen, a flow corresponds to a sequence of packets in which each packet has the same value for a 5-tuple of source and destination IP address, protocol number, and source and destination port number. This definition of flow makes it highly likely that all packets of a flow originate from the same application and the same network.

For a best representation of the header fields as well as a way to analyze their behavior, we have developed a header field mapping. In this mapping, for some header fields, the values are simply copied from the packets; for others the mapped value represents the increment or decrement between consecutive packets into a flow, and finally for some of them which the distribution of values is highly skewed, we can replace the original value by a transformation or function of the values. In the next we describe this mapping.

Let P_i^m be the packet header of the i-th packet of a flow consisting of m packets. $P_i^m(j)$ is a selected header field of P_i^m. For each field $P_i^m(j)$, a function χ_j performs a mapping into an integer value $F_i^m(j)$:

$$F_i^m(j) = \chi_j(P_i^m(j)) \quad (1)$$

For each packet, let

$$F_i^m = (F_i^m(1), F_i^m(2), \ldots) \quad (2)$$

denote a vector of integers, where we include the selected fields. For the complete flow we can define:

$$P_m = (P_1^m, P_2^m, .., P_m^m) \quad (3)$$

and

$$F^m = (F_1^m, F_2^m, .., F_m^m) \quad (4)$$

Note that the vector F^m can be viewed as a numerical representation of the m packet headers, as we substitute some selected packet header fields by integers.

Generally, a measured trace consists of a large number of flows. For analysis purposes, it is useful to classify these flows into a small number of classes or clusters such that the components within a cluster are very similar to each other. Later, one member from each cluster may be selected to represent the class.

Using the flow mapping described in this section, in a high-speed link we can find potentially a large variety of F^m flows. To study this variety among flows, we employ a cluster analysis technique. Our clustering analysis basically consists of mapping each component (field) into an n-dimensional space and identifying components that are close to each other. Here n is the number of parameters. The closeness between two components is measured by defining a distance measure. The Euclidian distance is the most commonly used distance metric and is defined as:

$$d = \left(\{ \sum (x_{ik} - x_{jk})^2 \} \right)^{0,5} \quad (5)$$

From a set of flows, we calculate the Euclidian distance between the F^m vectors, equation (4), and the results are stored in a distance matrix of flows. Initially, each vector F^m represents a flow and a cluster. Later, we search the smallest element of the distance matrix. Let d_{rs} the distance between clusters r and s, be the smallest. We merge clusters r and s and also merge any other cluster pairs that have the same distance. We have used the Minimum Spanning Tree hierarchical clustering technique (JAIN, 1991), which starts with n clusters of one component (flow) each and successively joins to the nearest clusters until be reached a specific distance between the clusters. For each m (number of packets per flow), we apply, separately, the clustering method. After all, *Patterns of Flows* are generated from the clusters.

Starting from a real trace, we break it down into flows which have the same number of packets and map each field using our function, as shown in Figure 3. Below we start to justify the use of the three (n) parameters which are employed in order to realize the clustering.

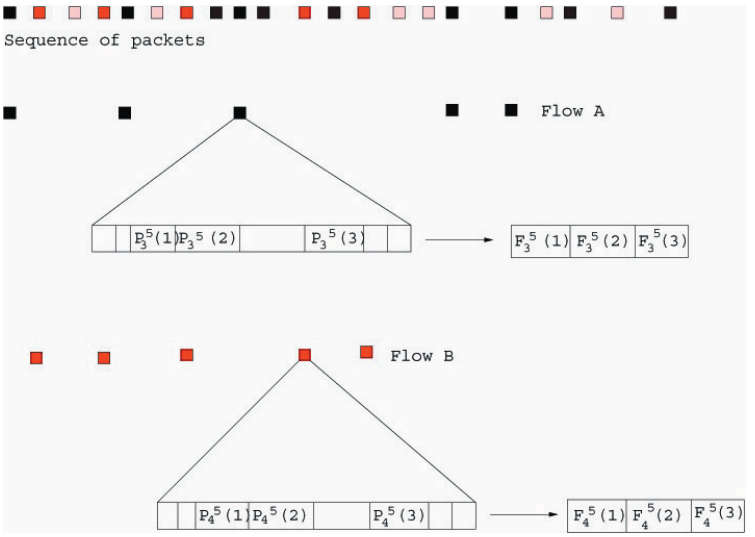

Figure 3: Flow Mapping

(i) TCP Flags: Improper use of TCP flags can be an important clue that an attack is taking place or is about to occur. Nowadays, Intrusion Detection Systems are capable of performing real-time traffic analysis looking for anomalies in traffic based on Connection-State-Machine and on TCP flags sequence. This includes unexpected TCP flags or invalid sequence numbers in received packets. Basically, we have used the Ack flags counts.

(ii) Total length: packet size distribution is a characteristic of network traffic that describes the size of packets traveling across the network. This distribution plays a significant role in the performance of networks. Observing the Internet packets, we have seen a huge range of values. These high values would dominate the cluster analysis. Thus, to utilize bytes transferred but dampen the effect of the high values we divided it by the total of packets.

(iii) Inter packet time into a flow: Analyzing a set of flows with the same number of packets we have noted that, for small flows, the behavior of the inter-packet time Δt between consecutive packet is very similar for the majority of flows. Basically, this inter-packet time is very small or is near the *Round Trip Time (RTT).*This behavior is related to the TCP properties. The sequence of figures from 4 to 8 shows, for a set of 1,100 flows with 6 packets, how similar is the behavior of inter-packet time between consecutive packets into small flows.

Figure 4 exhibits the time between the first and the second packet, which corresponds to RTT (*Round Trip Time*) of each flow. This time is a consequence of the three-way handshake established by the TCP protocol and this figure shows, basically, three categories of RTT: small, medium and large. The inter packet time between the second and third packet, shown in Figure 5 is smaller than the RTT and near zero, implying that for flows with 6 packets, normally, the third packet is sent immediately after the second packet.

In Figure 6 for the time between the third and fourth packet, we see that some of them are near zero and the others are placed slightly in the right side of the RTT curve. In Figure 7 we can observe the same behavior. Figure 8 shows the behavior of the last packet, and we see a high concentration of flows with Δt near zero. We demonstrate with these graphs that different Internet flows exhibit high similarity.

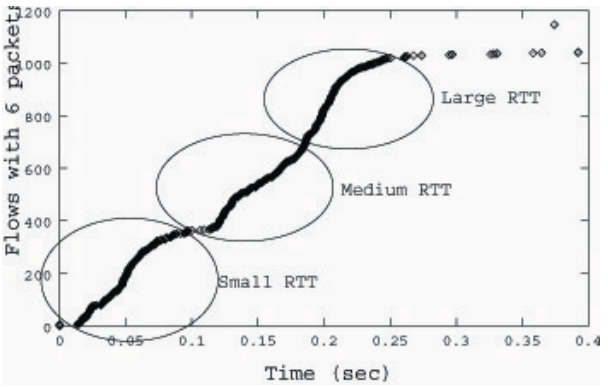

Figure 4: Δ t (1st and 2nd packets)

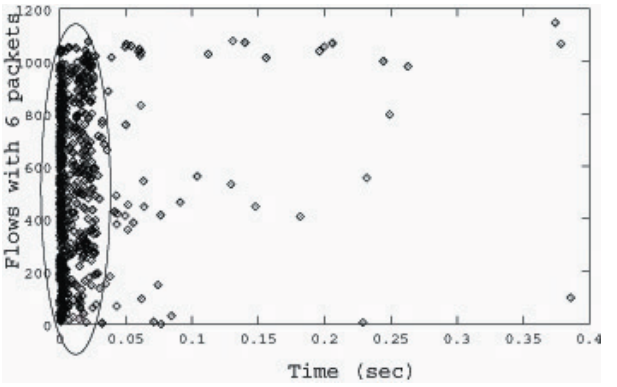

Figure 5: Δ t (2nd and 3nd packets)

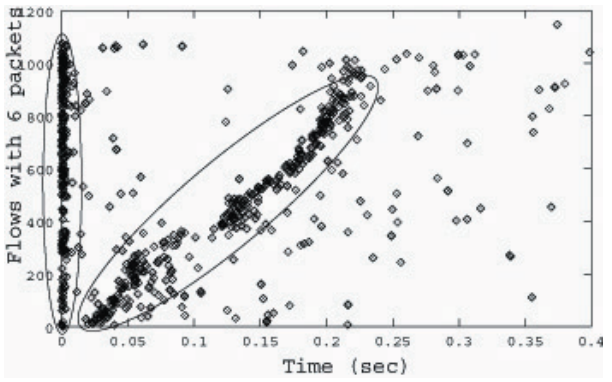

Figure 6: Δ t (3nd and 4nd packets)

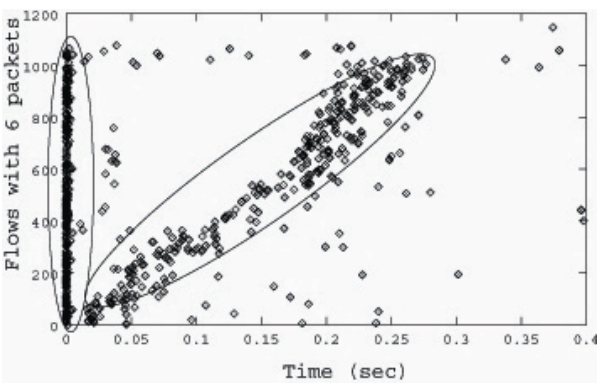

Figure 7: Δ t (4nd and 5nd packets)

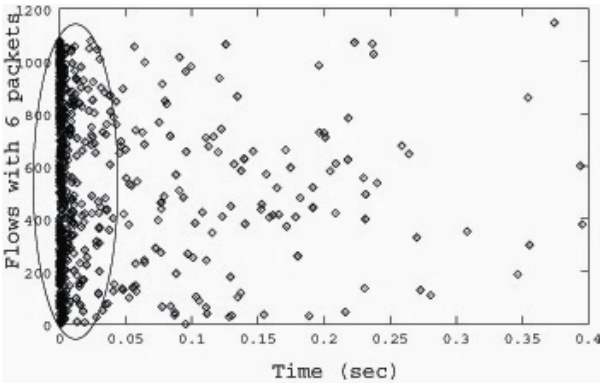

Figure 8: Δt (5nd and 6nd packets)

5 Packet trace classification

In this section we propose a methodology based on three steps to identify how similar are traces collected from different links and how different applications (e.g. WEB, P2P, FTP, etc) are distributed into the trace. The idea behind this proposed classification is to offer an easy and efficient method to select different patterns of traces to be used for performance evaluation purposes.

The first step is devoted to evaluate how distributed are packets among the m-packets flows. The Figure 9 shows, for four different traces, the percentage of packets (axis Y) placed in different m-packets flows (axis X). A first trace was collected on 1993 and has a high predominance of FTP traffic. The RedIRIS and Memphis University traces have a predominance of Web traffic but with the presence of P2P traffic. The last trace, the trace captured from Columbia University, shows a high predominance of Web traffic. Based only in this first step, we can see that two of them (RedIRIS and Memphis) show similar behaviour while the others have different distributions. From this first step, we have concluded that the trace collected on 1993 and the Columbia University trace are very different from the others.

The second step is devoted to study the variety among flows. The Figure 10 shows, for three traces and for m ranging from 2 to 13, the number of different clusters (axis Y) for each one of the m-packets flows (axis X). Taking the percentage of packets and the number of clusters per m-packets flows and applying a triangle-based cubic interpolation to create uniformly spaced grid data we display each trace (RedIRIS, Memphis University, Columbia University) as a surface plot (Figures 11, 12 and 13).

Looking at the shape of each figure, we can seee clearly that the RedIRIS and the Memphis traces show some similarities and that the Columbia trace shape is totally different.

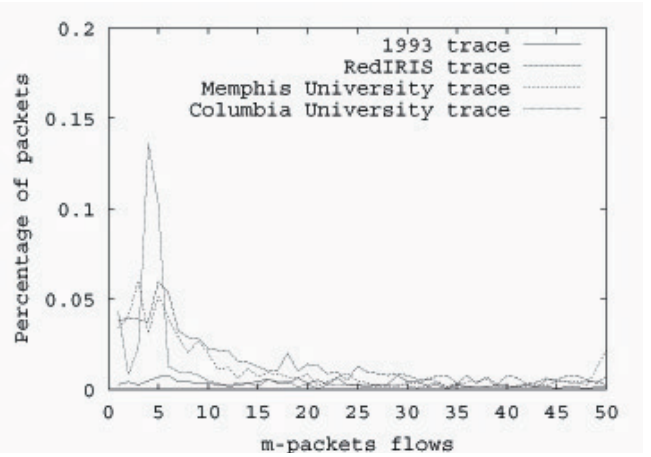

Figure 9: Packet Distribution

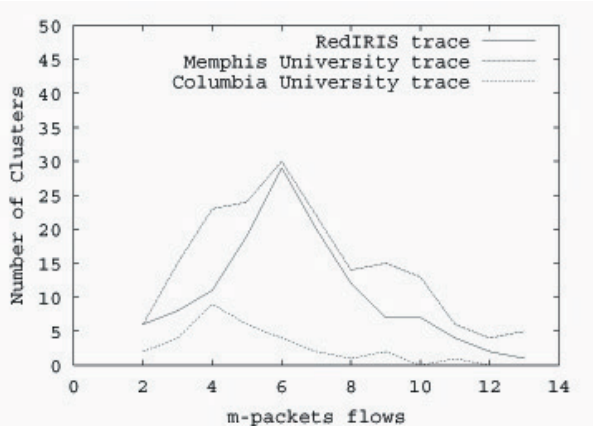

Figure 10: Number of Clusters

The third step analyze for each one of the m-packets flows, the amount of flows into each cluster. Using a flow clustering spectrum, we have represented on Figure 14, the spectrum for m=5. From each trace, we have selected the most representative clusters.

On Figure 14, each bar graph represents a trace and each colour on the bar graphic represents the percentage of flows that fit with this cluster. Plotting the spectrum of the three traces under analysis, we can see that the spectrum of RedIRIS and Memphis traces are similar while the Columbia trace shows a different spectrum. Similar outcomes were gathered for different m values.

After concluding these tree steps, we can be capable to identify with a high level of precision how semantically similar are different traces.

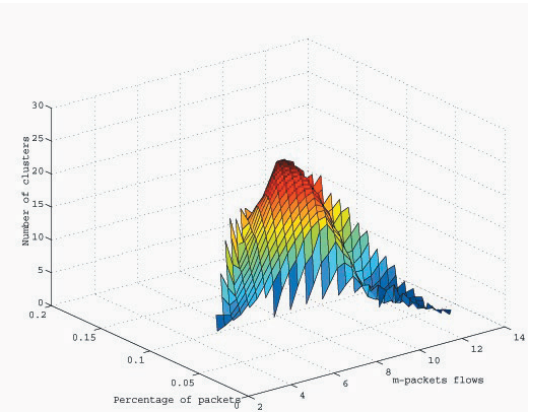

Figure 11: RedIRIS trace – 3D shaping

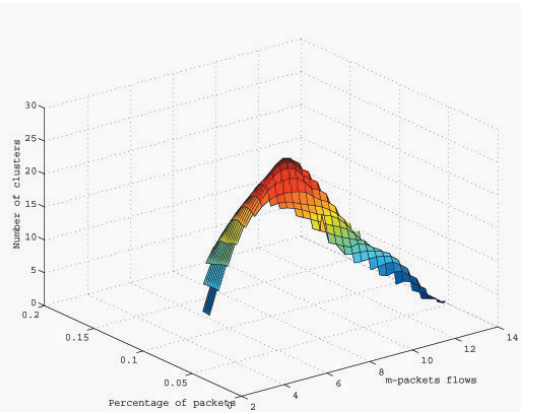

Figure 12: Memphis trace – 3D shaping

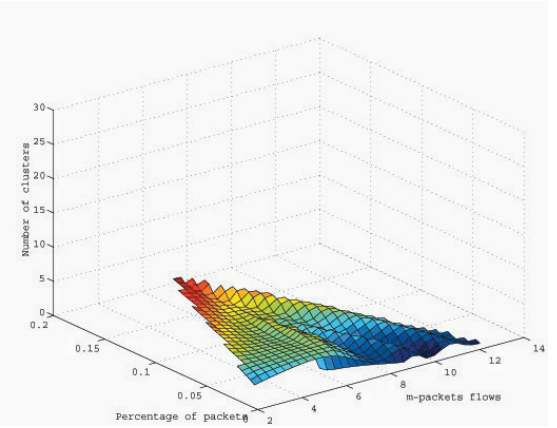

Figure 13: Memphis trace – 3D shaping

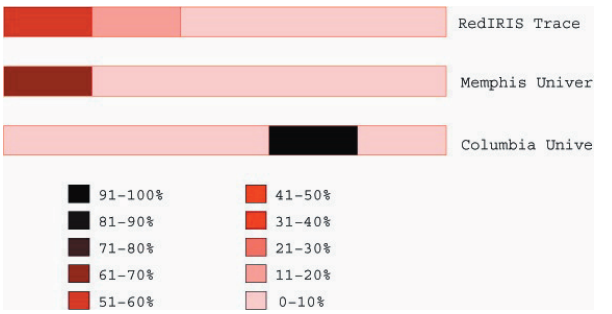

Figure 14: Flow clustering spectrum

6 Conclusions

This work investigated the use of the cluster analysis technique for identifying patterns of data flows on broadband traffic. That technique allows extractions of patterns of data flows. In the analysis and demonstration of the proposed method we employed TCP/IP traces from different sites to analyze and demonstrate the occurrence of flow clustering. Those traffics were modeled at different flow levels, each one composed by a 5-tuple of source and destination IP address, protocol number, and source and destination port number. This definition of flow makes it highly likely that all packets of a flow originate from the same application and the same network. In future works, we intend to extend the applicability of the packet trace classification method to identify different applications into any trace.

References

ALMEIDA V. et al. Characterizing reference locality in the WWW. In: INTERNATIONAL CONFERENCE ON PARALLEL AND DISTRIBUTED INFORMATION SYSTEMS PDIS96, 4., 1996, Florida. *Proceedings*… Florida: IEEE, 1996. p. 92-103. 1 CD-ROM.

BROWNLEE, N.; CLAFFY, K. Understanding Internet traffic streams: dragonflies and tortoises. *IEEE Communications Magazine*, New York, v. 40, p. 110-117, 2002.

GUO, L.; MATTA, I. *The war between mice and elephants*. Boston: Boston University, Computer Science Department, 2001. Technical Report BUCS-2001-005.

HOLANDA, R. A new methodology for packet trace classification and compression based on semantic traffic characterization. Ph.D Thesis, September, 2005, Catalunha, Catalunha University, 2005.

HOLANDA, R. et al. Performance analysis of a new packet trace compressor based on TCP flow clustering. In: INTERNATIONAL SYMPOSIUM ON PERFORMANCE ANALYSIS OF SYSTEMS AND SOFTWARE – IEEE/ ISPASS. 2005. *Proceedings*…Austin, 2005. 1 CD-ROM.

JAIN, R. *The art of computer systems performance analysis*. New York: John Wiley and Sons, 1991.

JOHNSON, D. E. *Applied multivariate methods for data analysis*. Minnesota: Brooks/Cole Publishing, 1998. 211 p.

KAUFMAN, L.; ROUSSEEUW, P. J. *Finding groups in data*: an introduction to cluster analysis. New York: Wiley and Sons, 1990.

KOHLER, E. et al. Observed structure of address in IP traffic. In: INTERNET MEASUREMENTS WORKSHOP - SIGCOMM, 2002, Marseille. *Proceedings*… Marseille, 2002. 1 CD-ROM.

LAKHINA, A.; CROVELLA, M.; DIOT,C. *Characterization of network-wide anomalies in trafic flows.* Boston: Boston University, 2004. Technical Report BUCS-TR-2004-020.

LAKHINA, A. et al. Structural analysis of network traffic flows. In: JOINT INTERNATIONAL CONFERENCE ON MEASUREMENT AND MODELING OF COMPUTER SYSTEMS-SIGMETRICS-PERFORMANCE'04. New York, 2004. 1 CD-ROM.

LEDYARD, T.; LEWIS, C. A reliability coefficient for maximum likelihood factor analysis. *Physicometrika Journal*, Grensboro, v. 38, p. 1-10, 1973.

MARCHETTE, D. A statistical method for profiling network traffic. In: WORKSHOP ON INTRUSION DETECTION AND NETWORK MONITORING, 1999, California. *Proceedings*... California, 1999. 1 CD-ROM.

NLANR. *Measurement and network analysis.* 1991.Disponível em: <http://moat.nlanr.net>. Acesso em 20 mar. 2005.

REDRIS. *Spanish national research network.* 1991. Disponível em: <http://www.rediris.es>. Acesso em: 20 mar. 2005.

TAYLOR, C. ; ALVES-FOSS, J. *An Empirical Analysis of NATE - Network Analysis of Anomalous Traffic Events*, In: NEW SECURITY PARADIGMS WORKSHOP, 10., 2002, Hampton. *Proceedings*…Hampton, 2002. 1 CD-ROM.

ABOUT THE AUTHORS

Raimir Holanda
Doctor in Computer Science (Technical University of Catalonia - Spain, 2005), Professor at University of Fortaleza – Computer Science Departament (MIA). Fortaleza - Ce, Brazil. E-mail: raimir@unifor.br

José Everardo Bessa Maia
Master in Eletrical Engeneering (University of Campinas - Brazil, 1985), Professor at University of Fortaleza; Professor at State University of Ceará - Computer Science Departament, Fortaleza – CE, Brazil. E-mail: jmaia@unifor.br

Gabriel Paulino
Master student at University of Fortaleza (MIA), Bachelor in Computer Science (UFC, 1995) and Mathematics (UECE, 2003); System Analist at Federal Bureau of Data Processing (SERPRO), Fortaleza-CE, Brazil. E-mail:gabrieljr@yahoo.com