Inter-Operator Traffic Differentiation based on Multiscale Analysis

Nelson Coelho, Paulo Salvador, António Nogueira
DETI, University of Aveiro/Instituto de Telecommunicações
Aveiro, Portugal
email: {nelsonmiguel, salvador, nogueira}@ua.pt

***Abstract*—Web 2.0 changed the interaction paradigm of Internet users, placing them in a more active role as both producers and consumers of digital contents. This concept has also triggered the appearance of social networks and cloud computing services, which have an increasing contribution to the total traffic amount. The increasing Internet complexity brings new challenges to network operators and managers, which need to understand new applications and know the exact properties of the generated traffic. The ability to accurately map traffic patterns to their corresponding application can be used to build efficient traffic and user profiles, which can be extremely helpful in several critical tasks like network resources optimization, service differentiation and personalization, network management and security. This paper proposes a classification approach that is able to accurately differentiate traffic flows in a core network and associate them with their underlying applications, allowing the construction of accurate traffic and user profiles. By performing a wavelet decomposition and analyzing the obtained scalograms, the captured traffic can be fully characterized in terms of its time and frequency components. As the different frequency components of the traffic are inferred, an appropriate communication profile characteristic of each application type can be defined. This way, it is possible to identify the distinct applications that are being used by the different connected clients and build useful user profiles.**

***Keywords - traffic identification; profiling; multi-scale analysis; wavelet transform*.**

I. INTRODUCTION

The ability to accurately build efficient traffic and user profiles has a crucial importance in many network operation and management tasks; it can be used to infer the most appropriate bandwidth and delay requirements for each user or group of users, allowing and optimized distribution of the network resources and improving the values of the Quality of Service (QoS) parameters; it will allow network managers to easily create groups of users requesting similar contents, easing the delivery of appropriate and related contents and services; security standards can be improved because it will allow the detection of users presenting illicit profiles or profiles including unknown applications, triggering alarms and providing counter-actions while allowing the remaining connected clients to experience better QoS levels.

This paper proposes a methodology for the creation of traffic profiles based on the classification of collected traffic flows, that is, based on their mapping to the corresponding generating applications. The proposed classification approach performs a wavelet decomposition at several scales of analysis; it is known that lower scales comprise low frequency events, which are typically created by user clicks and applications synchronization events; mid-range frequency components are related to the creation of Internet sessions; higher scales of analysis capture higher-frequency events, such as packet arrivals and packet bursts. So, by decomposing the traffic generated by different clients running diverse applications and analyzing it at various time scales, the methodology will be able to build a *multi-scale application profile* depicting the different frequency components that are characteristic of the mostly used applications.

Figure 1 represents the generic architecture of a traffic classification system with QoS support. Aggregated traffic is monitored by several network probes, which collect the necessary traffic flows at representative time periods. Using wavelet transforms, the most relevant components of the traffic flows are extracted, resulting in scalograms that are used to build accurate profiles for each one of the Internet applications whose traffic belongs to the aggregate. The profiles of the different applications are stored in a Profiles Database, which at bootstrap, only contains known profiles created in controlled environments or classified using deep-packet inspection (these are known as *training traces*). While capturing and classifying traffic, the different traffic profiles can be updated with the newly inferred profiles, after a validation process that may include payload inspection or human validation. The classifier associates the captured traffic to different service classes, characterized by several QoS parameters. All inferred/calculated data will also feed the User Profiling module, which is responsible for updating the different user profiles.

The efficiency of the proposed traffic classification methodology will be evaluated by applying it to aggregated inter-operator traffic, captured in the backbone network of a tier 1 Internet Service Provider (ISP). The results obtained show that the approach is efficient, being able to accurately differentiate Internet applications and, thus, having the potential to be the key component of a traffic and user profiling architecture.

The rest of the paper is organized as follows: Section II presents some of the most relevant related work on traffic classification and profiling; Section III provides some background on multi-scale analysis; Section IV describes the proposed classification methodology; Section V presents the traffic traces that are used to test the proposed methodology; Section VI presents and discusses the main results obtained and, finally, Section VII presents the main conclusions.

II. BACKGROUND ON TRAFFIC CLASSIFICATION AND PROFILING

Traffic classification efforts started by simple port-based identification approaches, where ports used by the different traffic

 ISBN: 978-1-61208-300-1

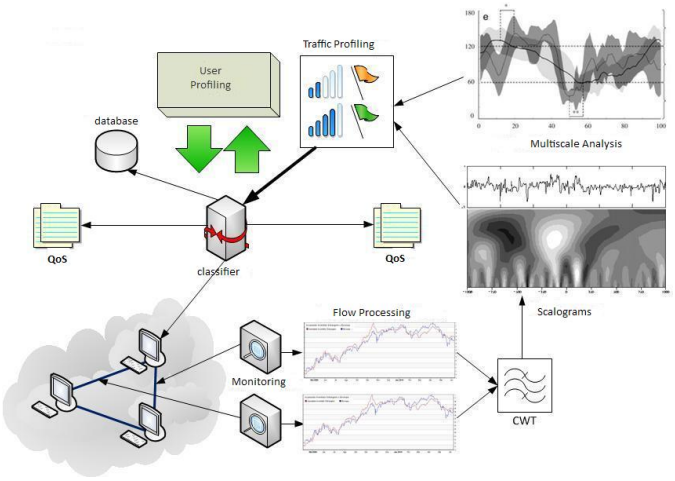

Fig. 1. Architecture of a traffic classification and user profiling system with QoS support.

flows were exclusively used to identify the applications that generated them. Since many protocols started to use random port numbers or ports generally associated to other protocols for bypassing *firewalls* and proxies, port-based approaches could no longer provide an accurate identification of Internet traffic [1].

Payload-inspection appeared as an evolutionary solution, inspecting the payload of captured packets in order to search for application level signatures of known applications. This approach relies on the use of extensive databases, containing known signatures and patterns of many Internet protocols, which are used as a comparison term whenever any new captured traffic has to be classified. This methodology was able to achieve good classification results, being used by several of the currently available commercial products [2] [3]. However, the databases associated to the classification approach need to be constantly updated in order to comply with new and emerging protocols. Besides, legal restrictions prevent Internet Service Providers from analyzing the contents of the users packets [4], while technical issues such as scalability on high-speed links also prevent researchers and Internet Service Providers from using payload inspection approaches.

Statistical analysis of the traffic flows appeared as the solution that could overcome these restrictions [5]. Moore *et al.* [6] proposed several flow discriminators and machine learning techniques to select the best discriminators for classifying flows. Hu *et al.* [7] built behavioral profiles describing dominant patterns of the studied applications and the classification results obtained showed that the approach was quite promising. In Huang *et al.* [8], authors attempted to describe negotiation behaviors by capturing traffic discriminators available at early negotiation stages of network flows and several machine learning algorithms were deployed to assess the classification accuracy. This way, they were able to conclude that the approach was suitable for *real-time* application identification. In a recent work [9], multi-dimensional probabilistic approaches were used to model the multi-scale traffic patterns generated by several Internet applications and to match the analyzed traffic to its generating application(s). However, these techniques can not efficiently differentiate between similar web-applications in scenarios where there is no access to layer 3 (and above) information and payloads. Hybrid classification approaches have also been used: in Tavallaee *et al.* [10], for example, a two-level hybrid approach in which payload analysis is combined with machine-learning algorithms was used to classify unknown traffic based on its statistical features.

There are several definitions of user profile [11], but a common definition can state that an user profile consists of a description of the user interests, behaviors and preferences. Therefore, the process of creating an user profile can be seen as the process of gathering the appropriate information until all these characteristics are obtained. In Claffy *et al.* [12], a parameterizable methodology for profiling Internet traffic flows at different granularities was proposed. Flows were defined based on traffic satisfying various temporal and spatial locality conditions, as observed at internal points of the network instead of only end-point definitions. In Xu *et al.* [13], a real-time behavior profiling system for high-speed Internet links was proposed, using flow-level information from continuous packet or flow monitoring systems and relying on data mining and information-theoretic techniques to automatically discover significant events based on the communication patterns of end-hosts. Reverse Domain Name System lookups, which are used to determine the domain name associated with an Internet Protocol (IP) address, have also been used to provide a simple association between a domain and the services it is known to run. A similar work was carried out in Trestian *et al.* [14], where authors stated that all information needed to profile any Internet endpoint is available in the Internet itself: therefore, accurate profiles were built by simply querying the most used search engine (Google) and dividing the querying results into several tags describing the requested services. However, inspection techniques can not be applied in scenarios where layer 3 and layer 4 information is not available, such as networks where authentication and encryption mechanisms are deployed.

III. Multi-Scale Analysis

The use of a wavelet decomposition through the Continuous Wavelet Transform (CWT) allows the analysis of any process in both time and frequency domains, being widely used in many different fields such as image analysis, data compression and traffic analysis. The CWT of a process $x(t)$ can be defined as [15]:

$$\Psi_x^{\psi}(\tau, s) = \frac{1}{\sqrt{|s|}} \int_{+\infty}^{-\infty} x(t)\psi^*(\frac{t-\tau}{s})dt \qquad (1)$$

where $*$ denotes the complex conjugation, $\frac{1}{\sqrt{|s|}}$ is used as an energy preservation factor, $\psi(t)$ is the *mother wavelet*, while τ and s are the translation and scale parameters, respectively. By varying these parameters, a multi-scale analysis of the entire captured process can be performed, providing a description of the different frequency components present in the decomposed process together with the time-intervals where each one of those components is located.

 ISBN: 978-1-61208-300-1

The wavelet scalogram can be defined as the normalized energy $\hat{E}_x(\tau, s)$ over all possible translations (set $\mathbf{T}$) in all analyzed scales (set $\mathbf{S}$), and is computed as:

$$\hat{E}_x(\tau, s) = 100 \frac{\left|\Psi_x^{\psi}(\tau, s)\right|^2}{\sum_{\tau' \in \mathbf{T}} \sum_{s' \in \mathbf{S}} \left|\Psi_x^{\psi}(\tau', s')\right|^2} \quad (2)$$

The volume bounded by the surface of the scalogram is the mean square value of the process. The analysis of these scalograms enables the discovery of the different frequency components, for each scale (frequency) of analysis. Assuming that the process $x(t)$ is stationary over time, several statistical metrics can be obtained, such as the standard deviation:

$$\sigma_{x,s} = \sqrt{\frac{1}{|\mathbf{T}|} \sum_{\tau \in \mathbf{T}} (\hat{E}_x(\tau, s) - \mu_{x,s})}, \forall s \in \mathbf{S} \quad (3)$$

where $\mu_{x,s} = \frac{1}{|\mathbf{T}|} \sum_{\tau \in \mathbf{T}} \hat{E}_x(\tau, s)$, and $|\mathbf{T}|$ denotes the cardinality of set $\mathbf{T}$.

IV. Classification Methodology

By defining characteristic regions of the scalogram statistics, for the different applications, in different frequency sub-sets, it is possible to identify profiles presenting components characteristic to each one of the applications. Such regions are inferred from the scalograms obtained from the decomposition of the *training traces* of each web-application.

Let us consider the (positive) region R_a^+ as the region defined as a function of a frequencies (positive) sub-set $\mathbf{s}_a^+$ and energy variation (positive) sub-set $\mathbf{\Sigma}_a^+$ for which we always have the characteristic statistical values of application a. Moreover, let us define the (negative) region R_a^- as a function of a frequencies (negative) sub-set $\mathbf{s}_a^-$ and energy variation (negative) sub-set $\mathbf{\Sigma}_a^-$ for which we never have characteristic statistical values of application a.

$$R_a^+ = f(\mathbf{s}_a^+, \mathbf{\Sigma}_a^+) \quad (4)$$

$$R_a^- = f(\mathbf{s}_a^-, \mathbf{\Sigma}_a^+) \quad (5)$$

A traffic trace process $x(t)$ is classified as belonging to application a if for all scales belonging to sub-set $\mathbf{s}_a^+$ the energy standard deviation $\sigma_{x,s}$ belongs to region R_a^+ and, simultaneously, for all scales belonging to sub-set $\mathbf{s}_a^-$ the energy standard deviation $\sigma_{x,s}$ does not belong to region R_a^-:

$$C(x) = a \Leftarrow \forall s \in \mathbf{s}_a^+, \sigma_{x,s} \in R_a^+ \wedge \forall s \in \mathbf{s}_a^-, \sigma_{x,s} \notin R_a^- \quad (6)$$

The classification decision can be made as soon as all conditions are met. Note that, even if time $\mathbf{T}$ grows and allows more classification precision, decisions can nevertheless be made with small $\mathbf{T}$ sub-sets (short-time analysis and decision).

The inference of regions R_a^+ and R_a^- (defined by $\mathbf{s}_a^+, \mathbf{\Sigma}_a^+, \mathbf{s}_a^-, \mathbf{\Sigma}_a^-$) can be performed by solving the following optimization problem:

$$\max_{\mathbf{s}_a^+, \mathbf{\Sigma}_a^+, \mathbf{s}_a^-, \mathbf{\Sigma}_a^-} \left(\sum_{\forall i \in \mathbf{I}_a} C(i) == a \right) \wedge \min_{\mathbf{s}_a^+, \mathbf{\Sigma}_a^+, \mathbf{s}_a^-, \mathbf{\Sigma}_a^-} \left(\sum_{\forall i \notin \mathbf{I}_a} C(i) == a \right), \forall a \quad (7)$$

where $==$ represents a comparison function witch outputs 1 if both terms are equal and 0 if terms are different. $\mathbf{I}_a$ represents the subset of processes (known as) belonging to web-application a. This optimization problem was solved (not for the optimal solution) using exhaustive search. However, more advanced algorithms can be applied to find (sub)optimal solutions.

Several regions can be created, in the various frequency sub-sets, for each studied application a. The higher the number of regions of an application, the higher the ability to analyze the different frequency components and consequently, a more accurate traffic mapping can be achieved. An algorithm was used to automatically define such regions (obviously satisfying the above conditions) using known simple geometrical equations, such as ellipses.

V. Traffic Traces

The traffic traces used in the evaluation tests were obtained from the Cooperative Association for Internet Data Analysis (CAIDA) [16]. Traces of this repository have one hour duration and were collected every month in each one of the passive monitors managed by the organization [17]. Specifically, the datasets used in this paper were collected on July 21, 2011, after 12:59 PM, at the Equinix data center, located in Chicago, Illinois, which is connected to the backbone line of a level 1 ISP between Chicago and Seattle, Washington. The connection is bidirectional, so traces contain packets circulating in both directions. After data collection, CAIDA proceeded to data anonymization due to legal and privacy requirements [16]. Captures are divided into several ".cap" files whose duration varies between fifty and fifty nine seconds. In our evaluation studies, we have considered the traffic traces corresponding to the first five minutes, in both directions. Traffic flows are identified by the traditional 5-tuple definition (source IP address, destination IP address, source port, destination port and protocol). Note that both cities are located on different time zones, which helps explaining the differences observed on the traffic flows in both directions. All tests were made on a desktop computer equipped with an Intel Core i5 CPU 650 @ 3.20GHz x 4 processor and running the *Ubuntu* 12.04 LTS operating system. *Tshark* and *wireshark* [18] were used to calculate the relevant statistics of the traffic traces.

The following applications, all contributing to relevant percentages of current Internet traffic, were considered in this study: Hypertext Transfer Protocol (HTTP), Simple Mail Transfer Protocol (SMTP), Real Time Streaming Protocol (RTSP), Mobile Status Notification Protocol (MSNP) and XBOX Live. For each service, four contexts were analyzed: downstream and upstream traffic on the client side, downstream and upstream traffic on the server side. However, due to space restrictions, we will only present here results corresponding to the client side, although a similar analysis also applies to the other scenarios.

VI. Evaluation Results

As an illustrative example (similar plots were also made for all the other applications), Figure 2 shows the traffic volume over

 ISBN: 978-1-61208-300-1

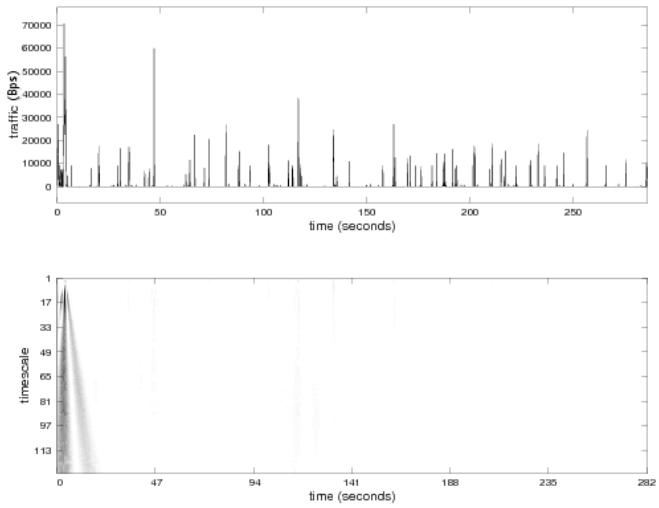

Fig. 2. HTTP downstream (from the client side) traffic: (up) traffic volume over time; (bottom) scalogram.

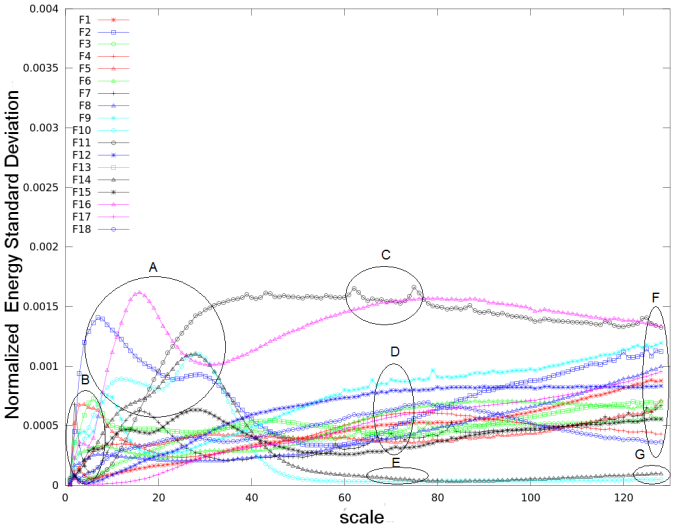

Fig. 3. Energy standard deviation of different HTTP downstream traffic flows.

time and the scalogram corresponding to HTTP downstream traffic. In this case, traffic peaks have low frequency, low amplitude and are not periodic, which indicates that this user is performing typical browsing. The corresponding scalogram exhibits visible frequency components at the beginning of the temporal scales range. Note, however, that HTTP traffic flows corresponding mainly to video downloads or file sharing have significantly different behaviors, so HTTP traffic presents a high diversity of patterns due to the wide spectrum of services that run on top of this protocol. This diversity is visible when analyzing flows from all the different HTTP usage profiles.

Figure 3 represents the energy standard deviation of different HTTP downstream traffic flows. Several regions can be identified, corresponding to different types of human and/or network events. Region A includes events having low frequency and moderate energy variation, usually generated by user clicks when accessing to online news, browsing photos and using social networks. Region B includes low frequency events with small energy variation, typical of the visualization of online video sites. Region C contains two flows (8 and 11) including medium frequency events with high energy variation, related to the creation of a high number of Transport Control Protocol (TCP) and HTTP sessions. Region D involves medium frequency events with a slower energy variation. Events included in region E have a very small energy variation, that is, a low number of TCP and HTTP sessions, typical of social networks applications (flows 10 and 14). Region F includes events with small and moderate energy variations, so flows located in this region have a moderate number of high frequency events, corresponding to the arrival of a reasonable number of packets. Finally, region G exhibits high frequency components with a very reduced energy variation, so flows 10 and 14 (which are located in this region) are generated by applications that are responsible for a reduced number of packets, like for example photo sharing or email clients.

In Figure 4, a single region is able to include all low frequency flows. The energy variation of these flows is small, indicating that they were generated by infrequent user clicks. In the medium frequencies range, we can find two flows (10 and 12), responsible for events with some energy variation (region B); region C includes the remaining flows, characterized by a small number of created sessions, which is responsible for their reduced energy variation. These characteristics are usually related to social networks applications and visualization of photos and videos. In the high frequency segment, two distinct regions can be identified: region D contains a flow (flow 12) characterized by a considerable percentage of high frequency events, which is a clear sign that this flow is responsible for significant upstream traffic originated at this client; region E incorporates the remaining flows and is characterized by events with high frequency components but with a reduced energy variation (much smaller when compared to region D), which corresponds to a reduced upstream packet rate characteristic of scenarios where users are not browsing in a very active way (which can correspond to video visualization scenarios, reading specific emails or reading news feeds on social networks' sites).

Figure 5 represents the energy standard deviation of different SMTP downstream traffic flows. In this case, most of the traffic flows have a similar behavior, being included in three different frequency regions: B, D and F. These flows have a small percentage of frequency components for all scales, which is associated to a low number of user clicks, few TCP sessions that were opened and a reduced number of packets that were exchanged. These results are according to the expected behavior of the client download traffic associated to SMTP. However, there are some flows diverging from this pattern. In region A (flows 2, 7, 11, 15 and 18) it is possible to find low frequency events with moderate energy variation, which implies more user clicks when comparing to flows located in region B. Region C (flows 7, 11, 15 and 18) incorporates medium frequency events with moderate energy variation; so, these flows have more HTTP and TCP interactions when compared to flows from region D. Finally, region E contains

 ISBN: 978-1-61208-300-1

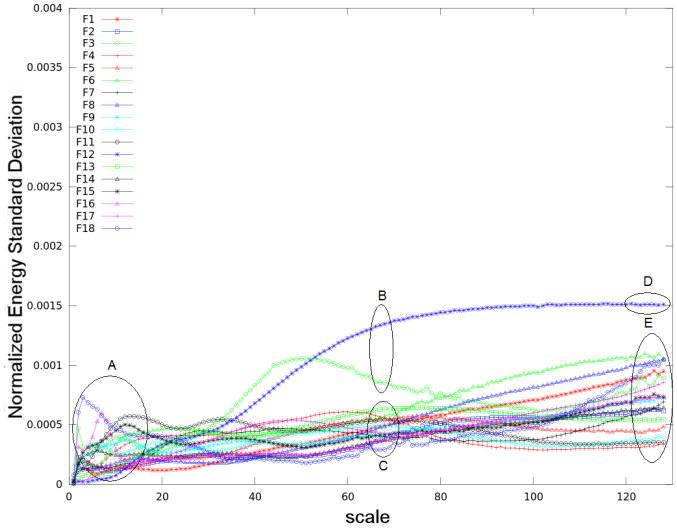

Fig. 4. Energy standard deviation of different HTTP upstream traffic flows.

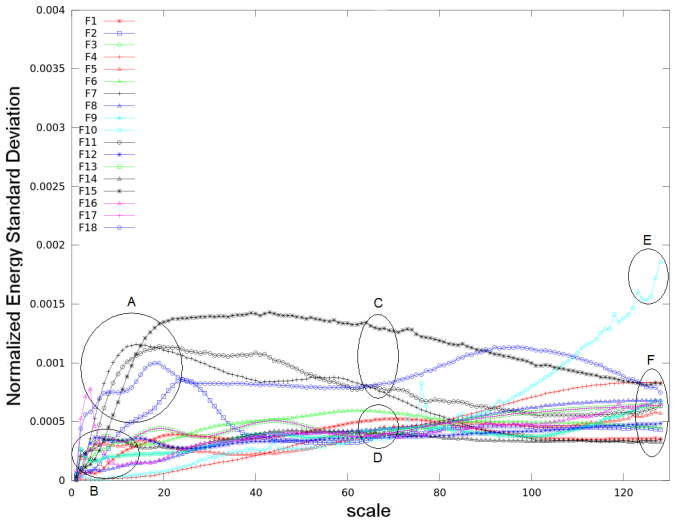

Fig. 5. Energy standard deviation of different SMTP downstream traffic flows.

a single flow (flow 10) with a very high energy variation, indicating that this flow corresponds to sending one or two large volume emails due to the high number of packets that were detected.

Figure 6, which represents the energy standard deviation of different SMTP upstream traffic flows, shows two regions in the low frequencies segment: region A includes very low frequency events, generated by the initial download of the email application interface and by the subsequent automatic synchronizations of the client email box. Region B corresponds to situations where, besides these events, the client performs other operations in his email interface, like sending emails to other contacts. Region C includes all traffic flows in the medium frequencies segment, having small to moderate energy variation, which correspond to the establishment of some TCP sessions during the analyzed time period. Finally, in the high frequencies segment two regions can be identified: region D, with a significant percentage of high frequency

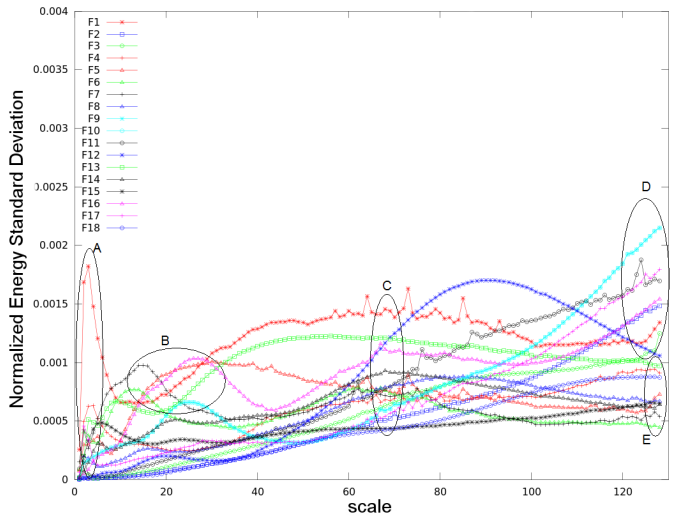

Fig. 6. Energy standard deviation of different SMTP upstream traffic flows.

components, and region E with a reduced percentage of high frequency components. Thus, region D is associated to events generated by sending large quantities of packets from the client to the server (specially, large emails), while region E is associated to situations where the client activity is more reduced and the packet exchange is only limited to control and synchronization data between the interface of the email application and the SMTP server.

Figure 7 represents the energy standard deviation of different RTSP downstream traffic flows. It is possible to verify the existence of several traffic flows corresponding to very low frequency events with high energy variation, which are generated by events such as automatic and periodic data synchronizations from the server and responses to clicks that were made by the client itself. Region B involves low frequency events with considerable energy variation, usually associated to requests for new contents; in the specific case of this protocol, this can correspond to the choice of new streams to visualize. In the medium frequencies segment, traffic flows are divided in two regions: region C includes flows with a considerable energy variation, which are obviously related to TCP and RTSP interactions; in fact, energy variations having this kind of amplitude should be related to client requests for establishing TCP and RTSP sessions dedicated to visualize streams. Region F includes high frequency events with a reduced percentage of high frequency components: the existence of few events in this region can be attributed to the reduced downstream packet transmission rate, which is quite unusual if streaming transmissions work as expected. So, we can assume that in this case the stream content was transmitted for a small period of time or some problems have occurred in the transmission while trying to visualize specific contents. Regarding region E, detected events have a considerable energy variation, which proves that the downstream packet transmission for these flows is higher than the one corresponding to flows of region F. Note that flows 8 and 9 have irregular peaks on the standard deviation of the normalized energy, specially in some specific intervals: region A, medium frequencies segment an very high

 ISBN: 978-1-61208-300-1

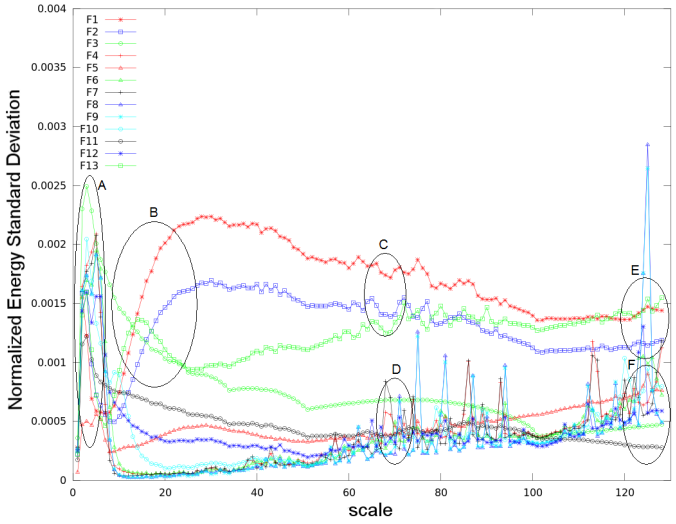

Fig. 7. Energy standard deviation of different RTSP downstream traffic flows.

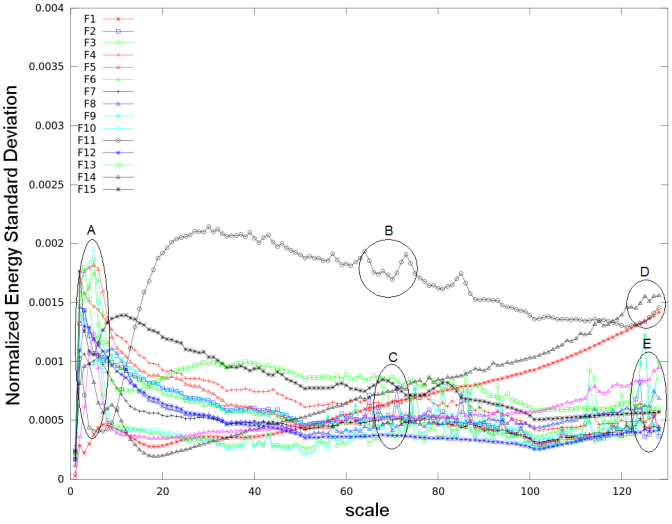

Fig. 8. Energy standard deviation of different RTSP upstream traffic flows.

frequencies segment. This irregularity in the energy pattern can be explained by the fluctuations on the packet transmission rate of video streaming applications: there are periods with high packet transmission rates (video is immediately processed and displayed), followed by periods where transmission rates are lower, resulting in some video freezing occurrences.

Figure 8 represents the energy standard deviation of different RTSP upstream traffic flows. Region A corresponds to very low frequency events, generated by rare occurrences: this should correspond to a very low number of client solicitations that are sent to the server or to the exchange of very small size packets. The medium frequencies segment contains two regions (B and C), although region B only contains one flow. This flow consists of events presenting energy variations higher than those corresponding to events from region C, so for this specific traffic flow there are more TCP and RTSP interactions and more TCP sessions established upon client request. We can assume that the client is trying to access various streams (in order to select the one that presents the highest quality) or is forced to update his browser because the quality of the connection/visualization is not as good as it should be. Regarding the high frequencies segment, we can observe that most of the analyzed traffic flows are located in region E and have small energy variation. The three flows located in region D correspond to events having higher energy variation than the one corresponding to events associated to region E, which means that there is a higher upstream packet transmission rate. Note that flows 3 and 9 have an irregular shape in the medium and very high frequencies segments, probably due to the instability of the connection between client and server.

Figure 9 represents the energy standard deviation of different MSNP downstream traffic flows. In the low frequencies segment, there is only one region, region A, including all traffic flows. Events corresponding to this region have a very low frequency, typically presenting a periodic pattern that makes sense if the time intervals between writing and reading messages is similar. In the medium frequencies segment, all flows (except flow 4) belong to region C and present a small energy variation because sending text messages to other online users creates a small number of sessions and few User Datagram Protocol (UDP) and MSNP interactions. Flow 4 stands out from the others because its events show a considerable energy variation. This can be explained by the fact that this MSNP user is talking at the same time as other online users, resulting in a higher number of MSNP and UDP interactions. In the high frequencies segment, we can see that most of the analyzed flows belong to region E and present a reduced packet arrival rate, which is expected if we take into account that text messages that are usually sent by this application require small packet sizes. Flows 4 and 7 do not follow this rule, being located at region D, where the energy variation of the generated events is slightly higher than the one corresponding to flows from region E because the packet arrival rate itself is higher. This again can be explained by the fact that this MSN client has established new conversations with other users.

Figure 10 represents the energy standard deviation of different MSNP upstream traffic flows. Flows located in region A present the same behavior of the corresponding flows represented in region A of Figure 9, although in that case packets were originated at the server while now they are originated at the client. The medium frequencies segment is divided in two regions: region C includes traffic flows with small energy variation, while region B includes flows with moderate energy variation, corresponding to more MSNP and UDP interactions, that is, in these flows the client interacts with more users of the same application. In the high frequencies segment, we can see that all flows (except flow 10) are located in region E. Here, packet transmission rates are usually small, suggesting that the client interacts with few users of the same application. Regarding region D, it only contains flow 10: events associated to this flow generate a quite reasonable packet transmission rate from the client, meaning that this particular client interacts with others in a very active way.

Figure 11 represents the energy standard deviation of dif-

 ISBN: 978-1-61208-300-1

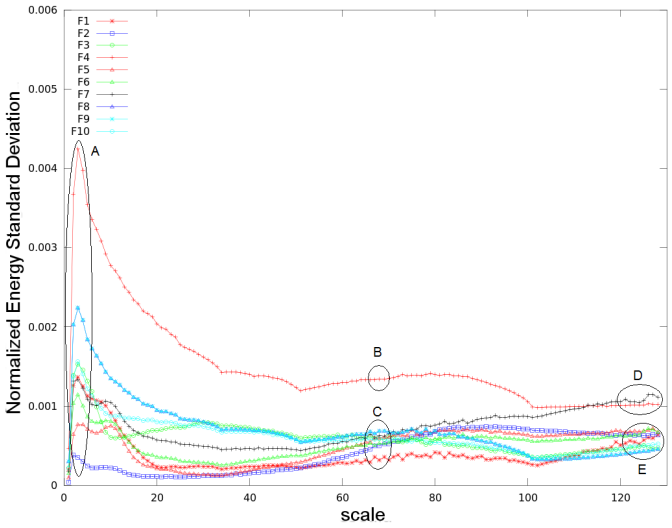

Fig. 9. Energy standard deviation of different MSNP downstream traffic flows.

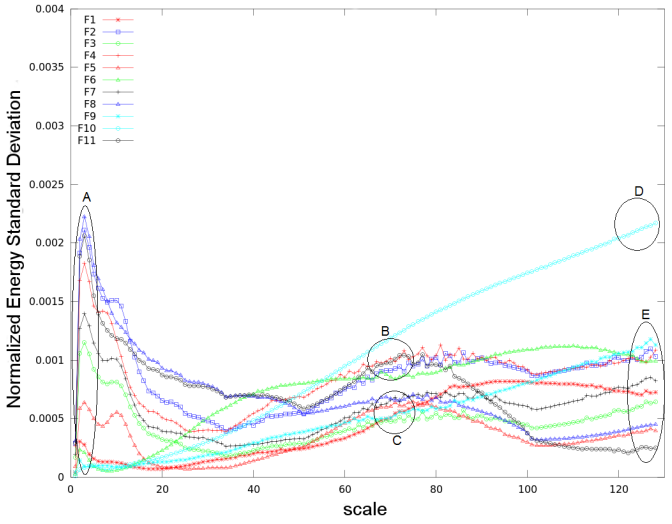

Fig. 10. Energy standard deviation of different MSNP upstream traffic flows.

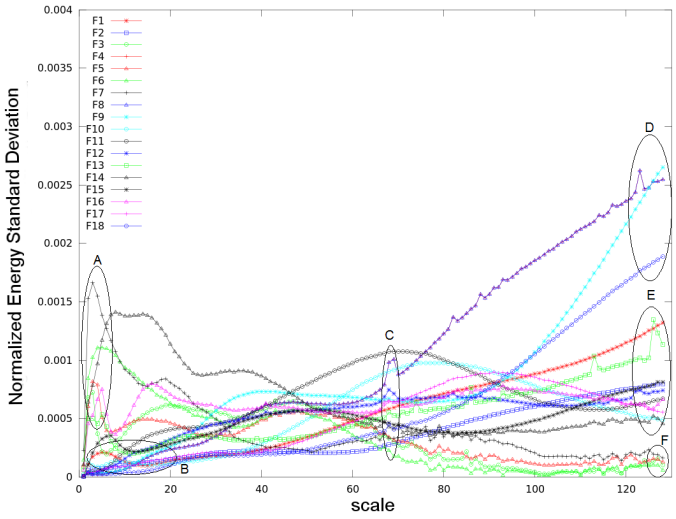

Fig. 11. Energy standard deviation of different XBOX downstream traffic flows.

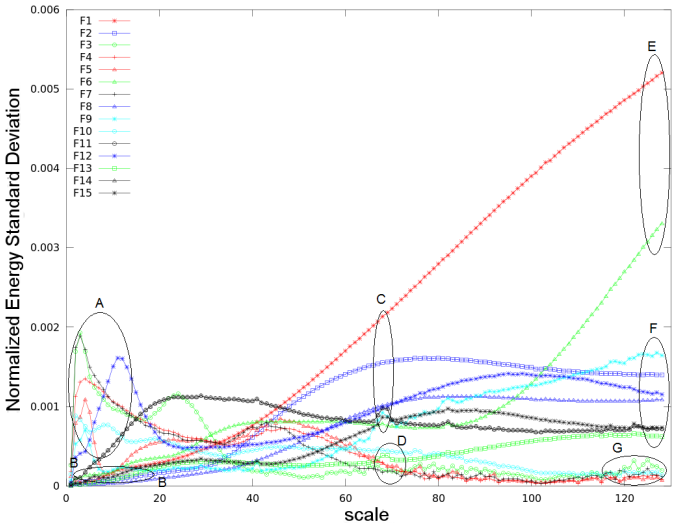

Fig. 12. Energy standard deviation of different XBOX upstream traffic flows.

ferent XBOX downstream traffic flows. There are two regions in the low frequencies segment: regions A and B. The first one includes very low frequency events, which rarely happen, such as the user activity in the XBOX Live service interface or automatic synchronizations between the remote server and the client terminal in order to assure an appropriate quality level for the connection. Region B includes low frequency events with small energy variation amplitude. Region C is located in the medium frequencies segment and includes all traffic flows because for this application energy variation does not differ significantly from one flow to another. The low value of energy variation indicates that few UDP sessions are created while the client is playing. In the high frequencies segment, three regions can be identified: region F, where the packet rate is very low; region E, where the packet rate is higher and region D where the packet reception rate is quite high (flows 4, 8, 9 and 18). It is possible to assume that flows of region D are associated to games where the user has a more active role, like personalized action games, sports or adventure games. Flows of region E can be associated to question/answer or to strategy games.

Finally, Figure 12 represents the energy standard deviation of different XBOX upstream traffic flows. Region E includes two flows (1 and 6) with high energy variation and high packet transmission rate between client and server, possibly corresponding to scenarios where the XBOX user is making a lot of clicks or playing a game that requires a high activity level. Region F, corresponding to a lower transmission rate, can be associated to games where the user is not so active and there are some inactivity periods between the different user actions (like, for example, games of the question/answer type).

From this discussion, we can conclude that a multi-scale analysis based on the wavelet transform is able to efficiently highlight the most important distinguishing features/characteristics of each Internet application. By performing a wavelet decomposition of the traffic flows, the different

 ISBN: 978-1-61208-300-1

TABLE I
PERCENTAGE OF CORRECTLY CLASSIFIED FLOWS.

Application	Download	Upload
HTTP	90.2%	93.1%
SMTP	77.8%	74.9%
RTSP	84.5%	82.3%
MSNP	91.1%	89.7%
XBOX	84.2%	88.6%

time and frequency components can be identified from the scalogram of the traffic metrics, allowing an efficient mapping of the captured traffic to the corresponding application.

Table I shows the classification results obtained by applying the previously discussed methodology. First of all, a set of *training flows* is used to define the elliptic areas that are used to differentiate the most relevant characteristics of each application; then, a set of *testing flows* is classified based on those areas in order to evaluate the efficiency of the classification approach. As can be seen, the classification results are quite good, with an identification accuracy higher than 70% for all cases. These classification results correspond to 50 runs and 95% confidence intervals were also built: since their widths are very small, they were omitted in the table.

VII. CONCLUSIONS

The complexity of current Internet forces network operators and managers to understand the underlying mechanisms of applications/services and the exact properties of the generated traffic. The ability to accurately map traffic patterns to their corresponding application can be used to build efficient traffic and user profiles that can be very useful in several operational and management tasks. This paper proposed a classification approach that is able to accurately differentiate traffic flows in the core network of a tier 1 ISP and associate them with their underlying applications. By performing a wavelet decomposition and analyzing the obtained scalograms, the captured traffic can be fully characterized in terms of its time and frequency components. This way, appropriate application profiles can be built, allowing the identification of all distinct applications that are being used by the different connected clients and the definition of useful user profiles.

REFERENCES

[1] A. Moore and K. Papagiannaki, “Toward the accurate identification of network applications,” Lecture Notes in Computer Science, vol. 3431, 2005, pp. 41–54.

[2] [retrieved: June, 2013] Snort home page. [Online]. Available: http://www.snort.org/

[3] [retrieved: June, 2013] Cisco IOS Intrusion Prevention System (IPS) - Products and Services. [Online]. Available: http://www.cisco.com/en/US/products/ps6634/index.html

[4] F. McSherry and R. Mahajan, “Differentially-private network trace analysis,” in Proceedings of ACM SIGCOMM, 2010, pp. 123–134.

[5] T. Nguyen and G. Armitage, “A survey of techniques for internet traffic classification using machine learning,” IEEE Communications Surveys Tutorials, vol. 10, 2008, no. 4, pp. 56–76.

[6] A. W. Moore and D. Zuev, “Internet traffic classification using bayesian analysis techniques,” in Proceedings of ACM SIGMETRICS, 2005, pp. 50–60.

[7] Y. Hu, D.-M. Chiu, and J. Lui, “Application identification based on network behavioral profiles,” 16th International Workshop on Quality of Service, 2008, pp. 219–228.

[8] N.-F. Huang, G.-Y. Jai, and H.-C. Chao, “Early identifying application traffic with application characteristics,” in IEEE International Conference on Communications, May 2008, pp. 5788 –5792.

[9] E. Rocha, P. Salvador, and A. Nogueira, “Detection of illicit network activities based on multivariate gaussian fitting of multi-scale traffic characteristics,” in IEEE International Conference on Communications, Jun. 2011.

[10] M. Tavallaee, W. Lu, and A. Ghorbani, “Online classification of network flows,” in Seventh Annual Communication Networks and Services Research Conference, May 2009, pp. 78 –85.

[11] D. Godoy and A. Amandi, “User profiling in personal information agents: a survey,” Knowledge Engineering Review, vol. 20, Dec. 2005, no. 4, pp. 329–361.

[12] K. Claffy, H. Braun, and G. Polyzos, “A parameterizable methodology for internet traffic flow profiling,” IEEE Journal of Selected Areas in Communications, vol. 13, Oct. 1995, no. 8, pp. 1481–1494.

[13] K. Xu, F. Wang, S. Bhattacharyya, and Z.-L. Zhang, “A real-time network traffic profiling system,” in 37th Annual IEEE/IFIP International Conference on Dependable Systems and Networks, Jun. 2007, pp. 595 –605.

[14] I. Trestian, S. Ranjan, A. Kuzmanovic, and A. Nucci, “Googling the internet: profiling internet endpoints via the world wide web,” IEEE/ACM Transactions on Networking, vol. 18, Apr. 2010, no. 2, pp. 666 –679.

[15] J. Slavic, I. Simonovski, and M. Boltezar, “Damping identification using a continuous wavelet transform: application to real data,” Journal of Sound and Vibration, vol. 262, 2003, no. 2, pp. 291 – 307.

[16] [retrieved: June, 2013] Caida - The Cooperative Association for Internet Data Analysis”. [Online]. Available: http://www.caida.org/home/

[17] M. Fomenkov and K. Claffy, “Internet measurement data management challenges,” in Workshop on Research Data Lifecycle Management, Princeton, NJ, Jul, 2011.

[18] [retrieved: June, 2013] Wireshark: go deep. [Online]. Available: http://www.wireshark.org/

 ISBN: 978-1-61208-300-1