Classification of Hidden Users' Profiles in Wireless Communications

Eduardo Rocha, Paulo Salvador, and António Nogueira

Instituto de Telecomunicações, University of Aveiro, Portugal
{eduardorocha,salvador,nogueira}@ua.pt

Abstract. The Internet can be seen as a mix of several services and applications running on top of common protocols. The emergence of several web-applications changed the users' interaction paradigm by placing them in a more active role allowing them to share photos, videos and much more. The analysis of the profile of each user, both in wired and wireless networks, becomes very interesting for tasks such as network resources optimization, service personalization and security. In this paper, we propose a promiscuous wireless passive monitoring classification approach that can accurately create users' profiles in terms of the used web-applications and does not require authentication with the wireless Access Point. By extracting appropriate layer 2 traffic metrics, performing a Wavelet Decomposition and analyzing the obtained scalograms, it is possible to analyze the traffic's time and frequency components. An appropriate communication profile can then be defined in order to describe this frequency spectrum which is characteristic to each web-based application. Consequently, it is possible to identify the applications that are being used by the different connected clients and build user-profiles. Wireless traffic generated by several connected clients running some of the most significant web-based applications was captured and analyzed and the obtained results show that it is possible to obtain an accurate application traffic mapping and an accurate user profiling.

Keywords: User Profiling, Web-Application Identification, Wireless Networks, Wavelet Decomposition, Scalograms.

1 Introduction

Nowadays, wireless networks are widely deployed and are an effective means for providing Internet connectivity over a large area to several users. In fact, with the increase on the usage of the Internet as the *de-facto* communications platform and with the emergence of mobile nodes and terminals with sophisticated connectivity capabilities, wireless broadband networks became the most used solutions for addressing all these issues. Among them, 802.11 networks are the most prevalent due to their ability to provide an high-bandwidth access in substantial coverage areas, together with an easy deployment. In such scenarios, the ability to accurately build efficient user-profiles can have a crucial importance for many different aspects. To begin with, one can more easily infer the

K. Pentikousis et al. (Eds.): MONAMI 2011, LNICST 97, pp. 3–16, 2012.

bandwidth and delay requirements that are more suitable for a certain user and network resources can then be optimized and better distributed among several users. Therefore, better Quality-of-Service (QoS) standards can be achieved by every connected client. Besides, by accurately profiling the connected users, network managers can create groups of users requesting similar contents, which eases the delivery of appropriate and related contents and services. In this way, revenues can be increased, while security can also be effectively improved since it is possible to detect users presenting illicit profiles or profiles presenting unknown applications, triggering alarms and providing counter-actions, such as disconnecting malicious users. In this manner, the remaining connected clients can experience a better quality of service and the network managers can make the best use of the network infrastructure.

In this paper, we propose a methodology for the creation of users profiles based on the analysis of used on-line web-based applications. Such analysis is achieved using a promiscuous wireless monitoring approach, being able to obtain an accurate profiling of the users that are connected to a given wireless network. This profiling approach does not require authentication with the Access Point (AP), being able to promiscuously monitor all connected clients and classify their hidden profiles. By collecting layer 2 traffic metrics, the proposed classification methodology performs a wavelet decomposition at several scales of analysis. In fact, it is known that lower scales of analysis comprise low frequency events, which are typically created by user clicks and applications synchronization events, while mid-range frequency components are related to the creation of Internet sessions. On the other hand, higher scales of analysis capture higher-frequency events, such as packet arrivals and packet bursts. By decomposing captured traffic generated by different clients running different web-based applications and analyzing it at the different scales, we build a *multi-scale application profile* which depicts the several frequency components characteristic to the most significant and used on-line web-based applications. As will be shown, these applications require and create different user interactions, thus creating different traffic patterns that lead to distinct frequency profiles. By analyzing such profiles and mapping their components into the corresponding user and/or network event, we can accurately map the captured traffic into its originating on-line web-based application. After inferring these characteristic profiles, classification can be performed as quickly as a perfect match is obtained. The speed of classification depends on the profile characteristics and can range from few seconds to few minutes, depending if differentiating characteristics appear at network/service scales or human scales, respectively.

The proposed profiling approach will be validated by analyzing traffic sent to several clients connected to a 802.11 wireless network and inferring the applications that are being run by the different clients. The obtained results prove that it is possible to accurately assign traffic to its originating on-line web-application, thus providing a reliable and accurate description on the usage of web-based applications. The use of Layer 2 metrics allows our classification approach to become appropriate for the classification of encrypted traffic, where the

packets' payloads are not available, and also to circumvent technological and legal restrictions that prevent the inspection of the packets contents.

The remaining part of this paper is organized as follows: Section 2 presents some of the most relevant related work on statistical classification of Internet traffic and behaviors; Section 3 presents some important background on wavelets and scalograms; Section 4 presents the system architecture and the classification methodology; Section 5 presents the validation of the proposed methodology by looking at the obtained results and, finally, Section 6 presents some brief conclusions about the conducted work.

2 Related Work

There are several definitions for an user profile [3], 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 in order to obtain all these characteristics. Many works, like for example [6], have addressed the issue of building accurate user-profiles describing the most important features, but the set of features and consequently the definition of the user profile vary according to the classification objective. In this paper, we adopt a very specific definition of user-profile, which is more oriented to the set of web-applications that each user runs and interacts with. Therefore, our work differs from the previously mentioned one in the fact that we describe an user profile as the set of used web-based applications, where the focus is placed on applications that allow users to share on-line information and contents.

Many approaches have been proposed to address the traffic classification problem, but classification methodologies themselves had to evolve with the sophistication and complexity of the Internet protocols. Indeed, classification approaches started by a simple port-based identification, where the ports used by the different traffic flows were unique identifying features of the applications that generated them. However, many protocols started to use random port numbers or ports generally associated to other protocols for bypassing *firewalls* and proxies and, therefore, port-based approaches could no longer provide an accurate identification of Internet traffic [8].

Payload-inspection appeared as an evolutionary approach, independent of the used ports, and consisted in inspecting the payload of the 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 allows an unequivocal classification of the captured traffic and many currently available commercial products deploy it [2] [1]. However, the databases associated to the classification approach need to be constantly updated in order to comply with new and emerging protocols. Besides, these port/payload inspection techniques can not be used to perform detailed web-application identification because they

run on top of the HTTP protocol and consequently all the traffic will present typical HTTP digital signatures.

In [7], the authors analyzed only the TCP SYN, FIN and RST flags in order to obtain connection-level information about P2P traffic. In [13], 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.

Inspection techniques can not be applied in scenarios where layer 3 and layer 4 information is not available, like networks where authentication and encryption mechanisms are deployed for securing communications.

Statistical analysis of traffic flows appeared as a solution that could overcome these restrictions, since only the headers of the packets are analyzed [10]. The main concept of this approach is that traffic generated by the same protocol will present the same profile. In [9], several flow discriminators were proposed and machine learning techniques were used to select the best discriminators for classifying flows. In [4], the authors built behavioral profiles that described dominant patterns of the studied applications and the classification results obtained showed that this approach was quite promising. In [5], the authors attempt 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. By using such discriminators, the authors were able to conclude that the proposed approach is suitable for *real-time* application identification. In a recent work [11], multi-dimensional probabilistic approaches were used to model the multi-scale traffic patterns generated by several Internet applications and to match the analyzed traffic with its generating application. 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.

A more pragmatical and simpler approach can consist in performing reverse-DNS lookups in order to determine the domain name associated with the contacted IP address. Subsequently, a simple association between the obtained domain and the services it is known to run can be performed. A similar work was carried out in [14], where the authors state that all the information needed to profile any Internet endpoint is available around us - in the Internet. Therefore, in order to accurately profile the authors we simply have to query the most used search engine (Google) and divide the querying results into several tags describing the requested services. The obtained results proved that the approach is suitable for the proposed purpose, enabling even more accurate results than some of the state-of-the-art tools.

3 Multi-scale Analysis Based on Wavelet Scalograms

The use of a wavelet decomposition through the Continuous Wavelet Transform (CWT) allows the analysis of any process in both time and frequency domains. Therefore, this tool is widely used in many different fields such as image analysis,

data compression and, more recently, in traffic analysis. The CWT of a process $x(t)$ can be defined as [12]:

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

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. The first parameter is used for shifting the mother wavelet in time, while the second parameter controls the width of the window analysis and, consequently, the frequency that is being analyzed. 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. A 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:

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

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. For instance, the existence of a peak in the scalogram at a low frequency indicates the existence of a low-frequency component in the analyzed time-series while a peak in the scalogram at a high-frequency corresponds to an existing high-frequency component. In addition, assuming that the process $x(t)$ is stationary over time, several statistical information, such as the standard deviation, can be obtained:

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

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 $\mathbf{T}$.

Nevertheless, this analysis concept can be applied to non-stationary processes by performing time-variant statistical analysis and modeling.

4 Classification Methodology and System Architecture

As mentioned in section 1, our system consists of a promiscuous monitoring probe that is capturing traffic from all clients connected to the Access Point (AP) providing connectivity for a given wireless network. Figure 1 shows a diagram depicting the general architecture and the different components of the proposed

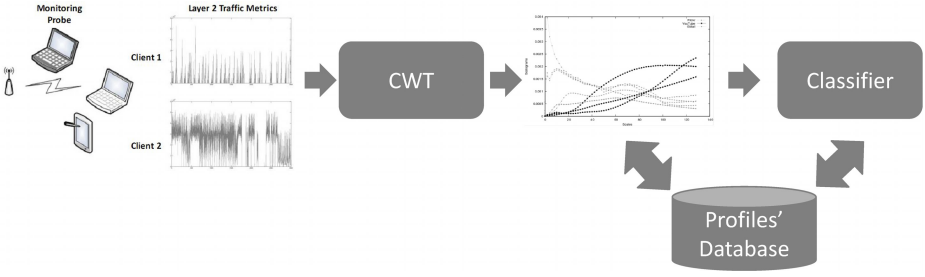

Fig. 1. Implementation Architecture

classification approach. To begin with, the wireless monitoring probe captures traffic sent to each one of the connected clients and, by extracting appropriate layer 2 metrics (such as the number of captured packets and bytes), it will be able to build a *traffic profile* of each connected node. The used monitoring probe does not require authentication with the Access Points (APs) of the wireless network and, therefore, can monitor nodes connected to several networks simultaneously. In addition, it can perform such tasks without being detected by any of the monitored clients, since the probe is not connected to the network.

The extracted layer 2 metrics are processed by a multi-scale analysis, enabled by a Continuous Wavelet Transform (CWT) decomposition, that builds a multi-scale traffic profile based on the corresponding scalogram that is inferred for some chosen statistical processes of the captured traffic traces, as presented in section 3. Such scalograms depict the several frequency components present in the captured traffic metrics, allowing the association of such components to the corresponding user/network event. For instance, low frequency components are related to low frequency events, which are usually created by user requests, and the scalogram allows the analysis of such events. Such requests create several Internet sessions which can be analyzed by inspecting mid-range frequencies components. Such Internet sessions lead to the download and upload of several network packets which consequently, create high-frequency components. By inspecting these different frequency components, one can infer the predominance, in the captured traffic, of the several presented user/network events, infer profiles characteristic to the different web-applications and classify the captured traffic. In fact, by analyzing statistical parameters such as the average, the standard deviation, or the correlation of the different scalograms, for each scale of analysis, we can infer the variability of the process energy and infer the most prominent frequency components. Such analysis will then assist us in finding differentiating frequency components in order to accurately classify traffic. The profiles of the several applications are stored in the Application Profiles' Database, which at bootstrap, contains only known profiles of the different applications (created in controlled environments or classified by deep-packet inspection). We refer to such traffic as *training traces*. While capturing and classifying traffic, the different profiles can be updated with the newly inferred frequency profiles, after a validation

process that may consist of payload inspection or human validation. In this way, an user profile can be obtained reflecting the most used web-applications.

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, we 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.

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

A traffic trace process $x(t)$ is classified as belonging to web-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^-:

$$C(x) = a \Leftarrow \forall s \in \mathbf{s}_a^+, \sigma_{x,s} \in R_a^+ \wedge \forall s \in \mathbf{s}_a^-, \sigma_{x,s} \notin R_a^- \tag{5}$$

The classification decision can be made as soon as all conditions are met. Note that, even if time $\mathbf{T}$ grows and allow 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:

$$\max_{\mathbf{s}_a^+, \mathbf{\Sigma}_a^+, \mathbf{s}_a^-, \mathbf{\Sigma}_a^-} \left(\sum_{\forall i \in \mathbf{I}_a} C(i) == a \right) \wedge \min_{\mathbf{s}_a^+, \mathbf{\Sigma}_a^+, \mathbf{s}_a^-, \mathbf{\Sigma}_a^-} \left(\sum_{\forall i \notin \mathbf{I}_a} C(i) == a \right), \forall a \tag{6}$$

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. Within the scope of this paper 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 several frequency sub-sets, for each studied web-application a. The higher the number of regions of an application, the higher the ability of analyzing the several frequency components and consequently, a more accurate traffic mapping can be achieved. An algorithm was created in order to automatically define such regions that satisfy the presented conditions by using known simple geometrical equations, such as ellipses.

5 Methodology Validation

In order to validate the proposed classification approach, several traffic measurements were performed as described in section 4. The analyzed traffic was collected by using a promiscuous monitoring probe that captures all traffic sent to each client connected to a 802.11 wireless network that was assembled at our networks laboratory. The different captured traffic flows were separated according to the destination MAC address, since our probe does not connect to the wireless network and does not access layer 3 traffic information. The layer 2 metrics considered for analysis were the number of captured bytes per sampling interval (0.1 seconds).

Five significant on-line Internet services were considered for analysis: on-line news, on-line mail, social networking, photo sharing and video services. Several usage scenarios were created to generate traffic from these services: for example, on-line news traffic was generated by visiting the most important Portuguese newspaper site (www.publico.pt) and browsing through the available news; on-line video download traffic was generated by watching videos in YouTube; for generating traffic from an on-line photo-sharing application, an account was created in Flickr (www.flickr.com) and only the traffic generated while browsing other users' photos was considered for analysis; on-line e-mail traffic was generated by using the services offered by GMail, specifically traffic generated only by the automatic synchronizations between the client web-terminal and the GMail server; finally, social networking traffic was generated by using an account created on Facebook (www.facebook.com) and interacting with the news updates coming from the remaining connected users, which does not include chatting and gaming. Table 1 shows the mapping between the available web-applications and the web sites that were used to generate traffic from each service.

5.1 Analysis of the Traffic Scalogram

As mentioned in sections 1 and 4, a wavelet decomposition of the captured layer 2 traffic metrics was performed using the CWT. The obtained scalograms were normalized for the whole length of the process, as described in equation 2. Figures 2 to 6 show the captured traffic metrics, the download rate in bytes per second, sampled in 0.1 seconds intervals, and the corresponding wavelet scalograms to the different web-applications that were mentioned in section 3. The analysis of these figures reveals differentiating characteristics that are caused by the distinct traffic patterns presented by these applications that have origin in human and network/service interaction characteristics. On-line news traffic (Figure 2), for example, presents several aperiodic peaks of short duration and considerable amplitude. Such peaks are caused by the user clicks on hyper links while browsing through the available news, causing the download of a new page that presents the requested news, creating considerable low frequency components. In addition, the scalograms generated by this application present some considerable mid-frequency components, due to the considerable number of created Transmission Control Protocol (TCP) sessions, while there are some

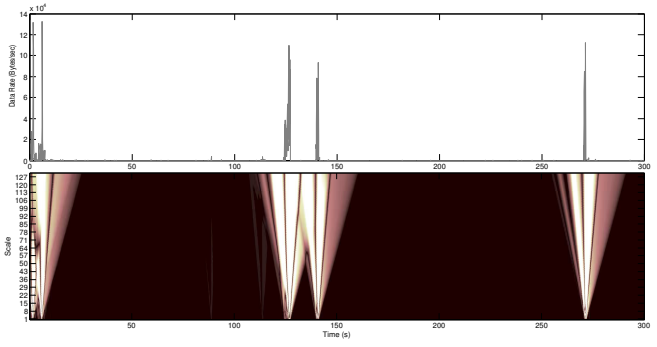

Fig. 2. On-Line News Traffic Patterns and corresponding Wavelet Scalograms

considerable high frequency components due to packets arrivals. On-Line video services (Figure 3) generate high-bandwidth traffic with a low Inter-Arrival Time (IAT) between packets, which is caused by the download of the requested video at the full available network bandwidth. Consequently, there are considerable high-frequency components, caused by packets arrivals, while there are no considerable low-frequency components since there are not so many user clicks. On-line Photo-sharing (Figure 4) applications usually generate several traffic peaks with pseudo-periodicity, due to the clicks that are performed by the user while requesting to see another picture. Such peaks are usually of low amplitude, since they only consist on the download of one picture using a single TCP session. Consequently, we can notice several high frequency components, of low amplitude, spread over the corresponding scalogram, while there are also are some low frequency components. On-line email applications (Figure 5) generate traffic presenting very low frequent traffic peaks, corresponding to the initial and automatic synchronization between server and client. These peaks are of very short duration and are less frequent than the ones of the previously presented online applications. Therefore, there are small high-frequency components caused by the synchronization traffic that merely checks for new e-mails, while low-frequency components are not very spread over the traffic scalogram due to near periodical nature of network/service events. Finally, on-line social networking applications (Figure 6) generate traffic presenting more frequent traffic peaks, of lower amplitude, which are generated by the status updates created by other connected users, which usually consist only of text messages. Therefore, there are less low-frequency components, while the high-frequency components are also less present in the process due to the small amount of traffic exchanged.

Figure 7 presents a graph of the standard deviation (over time) of the wavelet coefficients versus the corresponding frequency, or scale of analysis, of four different flows (randomly chosen from the data-set) belonging to each web-application. According to this figure, by analyzing the variation profile of the network process energy throughout the whole range of frequencies it is possible to obtain an accurate association between a given traffic flow and the application that originated it, by performing an analysis in the differentiating regions as explained in

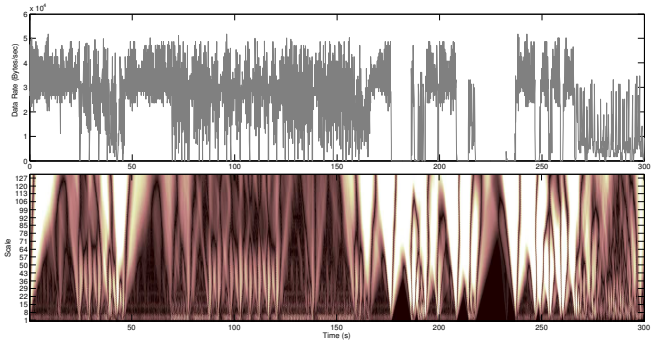

Fig. 3. On-Line Video Traffic Patterns and corresponding Wavelet Scalograms

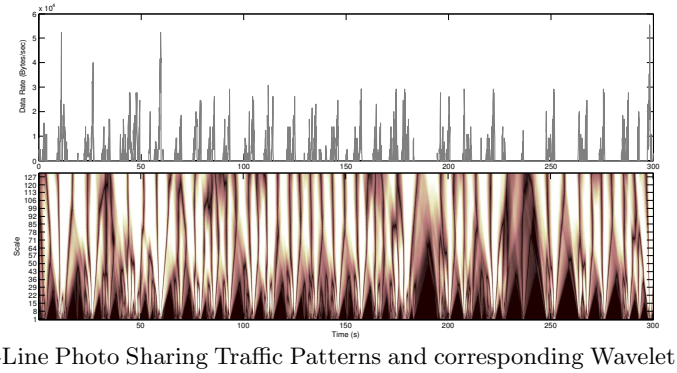

Fig. 4. On-Line Photo Sharing Traffic Patterns and corresponding Wavelet Scalograms

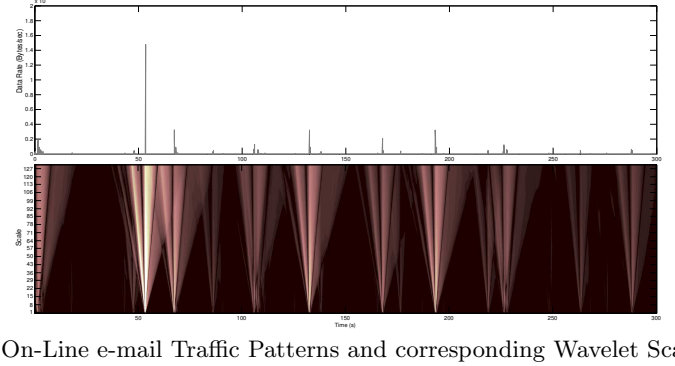

Fig. 5. On-Line e-mail Traffic Patterns and corresponding Wavelet Scalograms

section 4. The depicted regions were inferred by solving the minimization processes described in equation (6) using exhaustive search algorithms in predefined solution sets using the complete dataset.

Let us begin by analyzing the inferred regions and describing the differentiating traffic characteristics that led to them, since each region characterizes a subfrequency range that is mapped into specific human and network/service events.

For instance, region A comprises very-low frequency events, usually triggered by very rare events. This region, as shown in figure 7, encompasses traffic from on-line e-mail, mostly generated by the initial download of the e-mail web interface. Region B encompasses low frequency events, such as user clicks requesting new contents suitable to on-line news browsing or browsing through pictures on an on-line photo-sharing community or interacting in social networking applications. Therefore, the differentiation between these three applications will have to include more mid and high frequency regions. Regions D and E encompass mid-frequency events such as the ones associated with TCP and HTTP interactions, but the first region includes traffic presenting higher energy variation in this frequency range, implying that a higher number of TCP sessions are created: this behavior is more likely to be created by user clicks on on-line news sites, since the download of a new page comprises several TCP and HTTP sessions. On the other hand, the second region (E) includes traffic presenting a lower number of created sessions, since there is lower energy variation in that region of frequencies. This is more characteristic of social-networking applications, since the interaction with the news feed and the corresponding status updates create less TCP sessions than the previously mentioned application. Region C includes traffic that presents a low energy variation on low-frequency events, such as user clicks, or events with similar inter-event time. Both characteristics can be associated to on-line video applications, since they require a low number of user clicks, and photo-sharing applications, where the time between clicks presents lower variation. Region F is more characterized by a significant amount of high frequency events, such as packets arrivals, suitable to describe the high-frequency profile created by on-line video applications or web-pages with embedded video, characteristic of on-line news applications. On the other hand, region H can be seen as a region that is more characteristic to applications with a low number of events on such frequencies, which is suitable to describe Internet applications like photo-sharing since a low number of packets is required to download a shared picture. Region G is located between the two previously mentioned regions and presents more significant high-frequency components than region H and less high-frequency components than region F. Such region can be used to identify flows with a considerable (but not high) packet arrival rate. Therefore, each studied web-application can be mapped into one or more of the presented regions, as shown in table 1, and an algorithm was created to detect and classify the scalograms of the different captured traffic flows. Such algorithm simply needs to detect the variation of frequency components of the several scalograms in the inferred regions, mapped into web-applications as described in table 1, and assign the corresponding traffic accordingly.

5.2 Classification Results

Let us now analyze the classification results that were achieved by applying the above presented approach and shown in table 2. One can conclude that most of the generated traffic is accurately mapped into the corresponding web-application. However, there are some classification errors that can be explained.

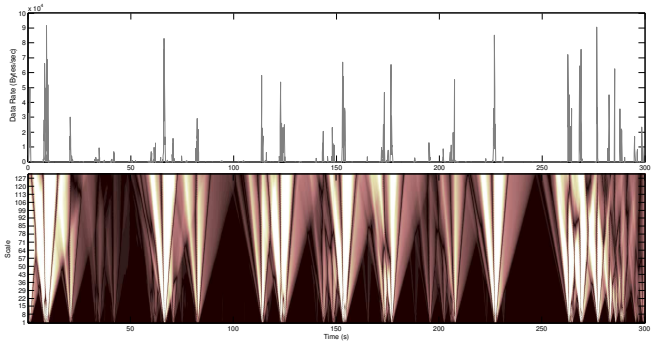

Fig. 6. On-Line Social Networking Traffic Patterns and corresponding Wavelet Scalograms

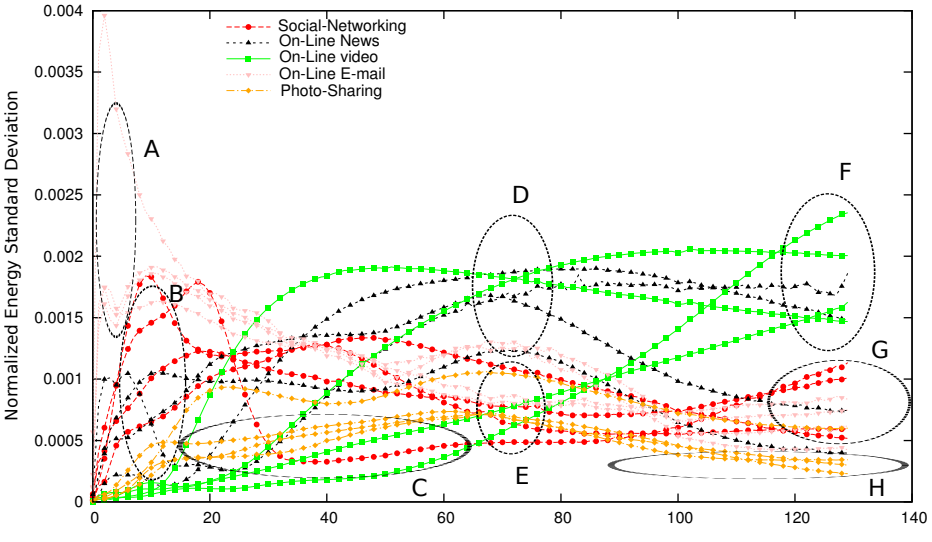

Fig. 7. Differentiating Regions

The association of some on-line traffic to video services can be due to the fact that some requested news presented embedded videos. Therefore, the profile can become similar to the one corresponding to video applications. Some flows from web-video traffic were assigned to on-line news, which can happen when watching several small duration movies since in this case the user can make more clicks in order to request for new contents, creating significant low-frequency components characteristic of on-line web-applications. Some classification mistakes also occurred for photo-sharing applications where some flows were classified as social-networking flows, which can occur when an Internet user visits the profile of another user connected in the same network. Some web e-mail was also associated to social networking applications, which can be due to the fact that

Table 1. On-Line Applications with their corresponding web sites and frequency mapping regions

Service	Web site	Regions
On-Line News	Publico (www.publico.pt)	B and D and (G or F)
On-Line Video	YouTube (www.youtube.com)	C and F and not(B)
Photo Sharing	Flickr (www.flickr.com)	B and E and H
On-Line E-mail	GMail (www.gmail.com)	A and (E or D)
Social Networking	Facebook (www.facebook.com)	B and E and G

Table 2. Classification Results

	Classified as				
Web-application	On-Line News	On-Line Video	Photo-Sharing	On-Line E-mail	Social Net.
On-Line News	**88%**	12%	0%	0%	0%
On-Line Video	11.1%	**88.9%**	0%	0%	0%
Photo-Sharing	0%	0%	**85.7%**	0%	11.3%
On-Line E-mail	0%	0%	0%	**87.5%**	12.5%
Social Networking	11.5%	0%	0%	0%	**88.5%**

when there is a small amount of e-mail updates the application profile gets more similar to social networking small message exchange. Finally, some social networking flows were associated to on-line news, which can occur if a considerable number of status updates occurs in a small time frame.

6 Conclusion

An accurate user profiling can be of crucial importance to several networking tasks, such as resources management, services personalization and security. In fact, by describing an user profile in terms of the web-applications that are used, one can easily and timely infer the bandwidth (and other network resources) demands, provide similar and related contents and also detect users with profiles presenting illicit or unknown patterns, activating the corresponding and needed network defense mechanisms. In this paper, we presented an approach that allows the identification of several web-applications used by different clients connected to a wireless network. By using a traffic monitoring and capturing probe, which does not require authentication, we were able to infer layer 2 traffic metrics and perform a Continuous Wavelet Decomposition in order to infer the corresponding traffic scalograms. By analyzing the frequency components present in these scalograms, it was possible to easily map each captured traffic flow into the corresponding web-application. The results achieved show that the proposed approach can accurately identify the different web-applications that were run by the connected clients.

Acknowledgments. This research was supported in part by Fundação para a Ciência e Tecnologia, grant SFRH/BD/33256/2007, and research projects PTDC/EIA-EIA/115988/2009 and PTDC/EEA-TEL/101880/2008.

References

1. Cisco ios intrusion prevention system (ips) - products and services (March 2011), http://www.cisco.com/en/US/products/ps6634/index.html
2. Snort: Home page (March 2011), http://www.snort.org/
3. Godoy, D., Amandi, A.: User profiling in personal information agents: a survey. Knowledge Engineering Review 20(4), 329–361 (2005)
4. Hu, Y., Chiu, D.M., Lui, J.: Application identification based on network behavioral profiles. In: 16th International Workshop on Quality of Service, IWQoS 2008, pp. 219–228 (2008)
5. Huang, N.F., Jai, G.Y., Chao, H.C.: Early identifying application traffic with application characteristics. In: IEEE International Conference on Communications, ICC 2008, pp. 5788–5792 (May 2008)
6. Iglesias, J.A., Angelov, P., Ledezma, A., Sanchis, A.: Creating evolving user behavior profiles automatically. IEEE Transactions on Knowledge and Data Engineering 99 (2011) (preprints)
7. Madhukar, A., Williamson, C.: A longitudinal study of p2p traffic classification. In: 14th IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, MASCOTS 2006, pp. 179–188 (September 2006)
8. Moore, A.W., Papagiannaki, K.: Toward the Accurate Identification of Network Applications. In: Dovrolis, C. (ed.) PAM 2005. LNCS, vol. 3431, pp. 41–54. Springer, Heidelberg (2005)
9. Moore, A.W., Zuev, D.: Internet traffic classification using bayesian analysis techniques. In: ACM SIGMETRICS, pp. 50–60 (2005)
10. Nguyen, T., Armitage, G.: A survey of techniques for internet traffic classification using machine learning. IEEE Communications Surveys Tutorials 10(4), 56–76 (2008)
11. Rocha, E., Salvador, P., Nogueira, A.: Detection of Illicit Network Activities based on Multivariate Gaussian Fitting of Multi-Scale Traffic Characteristics. In: IEEE International Conference on Communications, ICC 2011 (June 2011)
12. Slavic, J., Simonovski, I., Boltezar, M.: Damping identification using a continuous wavelet transform: application to real data. Journal of Sound and Vibration 262(2), 291–307 (2003)
13. Tavallaee, M., Lu, W., Ghorbani, A.A.: Online classification of network flows. In: Proceedings of the 2009 Seventh Annual Communication Networks and Services Research Conference, pp. 78–85. IEEE Computer Society, Washington, DC (2009)
14. Trestian, I., Ranjan, S., Kuzmanovic, A., Nucci, A.: Googling the internet: Profiling internet endpoints via the world wide web. IEEE/ACM Transactions on Networking 18(2), 666–679 (2010)