Gaussian Fitting of Multi-scale Traffic Properties for Discriminating IP Applications

Eduardo Rocha, Paulo Salvador and António Nogueira
University of Aveiro/Instituto de Telecomunicações
Aveiro, Portugal
e-mails: {eduardorocha, salvador, nogueira}@ua.pt

***Abstract*—In the last years, there has been an increasing need to accurately assign traffic to its originating application or protocol. Several new protocols and services have appeared, such as VoIP or file sharing, creating additional identification challenges due to their peculiar behaviors, such as the use of random ports or ports associated to other protocols. The number and variety of security vulnerabilities and attacks that are carried out over the Internet has also drastically increased in recent years. Besides, privacy and confidentiality are also growing concerns for Internet users: traffic encryption is becoming widely used and, therefore, access to the user payload is more and more difficult. Therefore, new identification methodologies that can be accurate when applied to different types of traffic and be able to operate in cyphered traffic scenarios are needed. In this paper, we present an identification methodology that relies on a multiscale analysis of the traffic flows, differentiating them based on the probability that their characteristic multiscale behavior estimators belong to specific probability distributions whose parameters are inferred from traffic flows of real applications. The classical concept of traffic flow was replaced by the definition of *data stream*, which consists of all traffic (in the upload or download directions) of a local IP address that is univocally identified by a numeric identifier. The results achieved so far show that the proposed methodology is able to accurately classify licit traffic and also identify some of the most common Internet security attacks. Besides, this approach can also circumvent some of the most important drawbacks of existing identification methodologies, namely their inability to work under strict confidentiality restriction scenarios.**

Keywords: Application identification, multiscale analysis, wavelets, licit and illicit applications.

I. INTRODUCTION

Classifying Internet traffic is a critical task for many areas, such as traffic engineering, Quality-of-Service (QoS), access control and security/intrusion detection. In recent years, the emergence of diversified and demanding applications made some of the mostly used classification methodologies (like port-based classification or payload inspection) inadequate. Besides, the number and diversity of attacks to hosts and services in the Internet increased in a dramatic way. Among these new threats, *botnets* are some of the most severe and dangerous [24], being responsible for some of the most stealth attacks, such as Distributed Denial-of-Service, Spam and phishing e-mails [4], [7], [6]. A *botnet* is a network of compromised computers under the control of a master, the *bot* master, which issues commands to the compromised hosts. Usually, these communications are encrypted, which poses a significant obstacle for Intrusion Detection Systems (IDSes). Moreover, the distributed nature of these attacks and the evolving (from centralized to distributed) structure of the botnets [16] also makes them extremely difficult to prevent.

This paper presents a new technique for identifying licit and illicit traffic flows based on the classification of different multi-scale behavior estimators. The classification methodology relies on the probability that these estimators belong to a Gaussian distribution whose parameters are inferred from traffic flows of the real applications. This approach presents several advantages over existing ones, namely its compliance with privacy issues since only packet headers at the IP and/or IP security protocols levels are analyzed. This work is an extension of a previous work [28] that also analyzed the multi-scale behavior of sampled flows generated by different applications using a kind of ”blind” clustering to classify the multi-scale coefficients' estimators. Here, we assume that these estimators follow a Gaussian distribution and use a probabilistic methodology to classify them, thus being able to discriminate their underlying generating applications. Besides the three widely used Internet applications that were also considered in [28] (web-browsing, video streaming and BitTorrent), we also include two of the most common attacks that are used by *botnets*: (i) *port scanning* and (ii) *snapshots* of the users' desktops. The classification results that have been already obtained show that the proposed approach is very promising, while being immune to some of the main disadvantages of current detection methodologies.

In order to be able to classify the different interactions that an application creates, which may consist of several sessions with different end-hosts/servers (and we strongly believe that the analysis of these interactions as a whole provides a deeper insight into how the applications behave and can assist in traffic discrimination), the restrictive classical definition of flow was replaced by the definition of *data stream*, which consists of all traffic (in the upload or download directions) of a local IP address that is univocally identified by a numeric identifier.

The remaining part of this paper is organized as follows: Section II presents some related work in the fields of traffic classification and attacks identification, Section III presents some background on wavelets and multiscale analysis, Section IV presents the details of the identification methodology; Section V presents some identification results that were already obtained in order to evaluate the efficiency of the proposed methodology and, finally, Section VI presents some brief

 ISBN: 978-1-61208-103-8

conclusions about the conducted work.

II. Related Work

The issue of traffic classification has been studied for many years and many techniques have been proposed to address this problem. In an early stage, traffic was classified according to the ports used for communication. However, this analysis became inaccurate when new protocols, such as BitTorrent or VoIP protocols, started to use random ports or ports associated to other applications. In fact, in a study conducted by [21], port-based techniques were unable to classify most of the network traffic that was generated by Peer-to-Peer (P2P) protocols. Payload analysis was one of the techniques proposed to overcome this limitation. It consists on the inspection of the packet's payload searching for characteristic signatures that can identify the generating protocol. A study carried out in [14] used this technique to identify P2P traffic and the results achieved were very accurate. In another work [30], digital signatures were also used to classify P2P traffic. The results achieved were very accurate and the authors proved that the proposed methodology can be effective in high-speed networks. However, in recent years, traffic encryption is becoming widely used to guarantee the confidentiality of the exchanged data in the Internet and, therefore, in these scenarios the packet payload is no longer accessible. Besides, when traffic is not encrypted the access to the packet's payload may not be allowed due to privacy restrictions.

Statistical analysis of traffic flows appeared as a solution that could overcome these restrictions, since only the headers of the packets are analyzed. The main concept of this approach is that traffic generated by the same protocol will present the same profile. Karagiannis *et al.* tried to identify P2P traffic based on a three-level analysis: social, functional and application levels. The accuracy of the obtained results was very high [15]. In another work [13], the authors built behavioral profiles that describe dominant patterns of the studied applications and the results showed that this approach was quite promising. In [21], the authors only analyzed the TCP SYN, FIN and RST flags in order to obtain connection-level information about P2P traffic. This technique has several inherent drawbacks: traffic presenting unknown behavior cannot be classified; when traffic is transported through a secure tunnel, the port numbers and the TCP flags may not be available and, consequently, classification is not possible.

In the last years, the number of security vulnerabilities and attacks increased at a dramatic rate [29]. *Botnets* have emerged and became one of the most dangerous threats to on-line security, being used for a wide variety of illegal activities such as DDoS, Spam, flooding attacks and exploit scanning, just to name some of them [22]. Besides, they are undetected by *anti-virus* software and IDSes [4]. Most IDSes, such as Snort [2], perform intrusion detection based on the recognition of signatures and known patterns from security attacks. This can constitute an accurate detection methodology, but these defense mechanisms cannot detect *zero-day* threats and attacks with unknown profiles [17]. Of course, IDSes can protect their networks by classifying any traffic pattern that deviates from an already known normal profile as an attack. Although this strategy could make them able to detect *zero-day* attacks, the detection accuracy would decrease since some of these "abnormal" profiles may be originated by legitimate user actions.

The structure of the botnets is also evolving, becoming more complex and distributed. For instance, the C&C infrastructure evolved from a centralized one, in which IRC protocols were used for communication, to a distributed one where P2P protocols and networks are used. Moreover, these communications can also be embedded in the HTTP protocol. Therefore, the detection of these networks is becoming more difficult and new methodologies are needed for their accurate detection.

Several studies have been conducted in order to collect, analyze and understand how *botnets* work: [5] studies the communications between the Command and Control (C&C) server and the infected machines; [25] analyzed the network behavior of spammers; [8] conducted several basic studies of *botnet* dynamics; [9] proposed to use DNS sink holing technique for *botnet* study and pointed out the global diurnal behavior of *botnets*; finally, [6] studied the relationship between *botnets* and scanning/spamming activities.

Based on this knowledge, different approaches have been proposed to solve the *botnet* detection problem: in [26], the authors used DNS-based black hole list counter-intelligence to find *botnet* members that generate spam; in [27], the authors proposed a system to detect malware (including *botnets*) by aggregating traffic that shares the same external destination, have a similar payload and involves internal hosts with similar OS platforms; [20] proposed a machine learning based approach for *botnet* detection using some general network-level traffic features of chat-like protocols, such as IRC; finally, [12] describes BotHunter, which is a passive *botnet* detection system that uses dialog correlation to associate IDS events to a user-defined *bot* infection dialog model.

III. Wavelets and Multiscale Analysis

A wavelet $\psi(t)$ can be defined as a pass-band function oscillating at a central frequency f_0. By performing a scaling change, which may consist of an expansion or a compression, and a temporal shift, we obtain $\psi_{j,k}(t) = 2^{-j/2}\psi(2^{-j}t - k)$, that is the oscillating central frequency moves to $2^{-j}f_0$ and the origin of the temporal reference to $2^j k$. Note that *j* represents the temporal scale, *k* represents the k^{th} coefficient corresponding to scale *j*, with j_0 being the larger time scale. Wavelet decomposition also uses a low-pass function, $\phi_{j_0,k}(t)$, known as scaling function, that can be scaled and temporarily shifted in a similar way to function $\psi_{j,k}(t)$. Therefore, the definition of the Discrete Wavelet Transform (DWT) of a stochastic process $X(t)$ is [11]:

$$X(t) = \sum_k c_X(j_0,k)\phi_{j_0,k}(t) + \sum_{j=j_0}^{\infty}\sum_k d_X(j,k)\psi_{j,k}(t) \quad (1)$$

where $c_X(j_0,k)$ are the scaling coefficients and $d_X(j,k)$ are the wavelet coefficients. The estimators for the first order moment of the wavelet coefficients can be defined as:

 ISBN: 978-1-61208-103-8

$$\mu_j = \frac{1}{n_j} \sum_{k=1}^{n_j} |d_X(j,k)| \quad (2)$$

where n_j is the number of coefficients to be analyzed at scale j. The scaling behavior of any stochastic process can then be studied by an analysis of the Logscale diagrams, which consist of logarithm plots of these estimators with the scales [3].

As mentioned in Section I, phenomena such as Short-Range Dependence (SRD) and Long-Range Dependence (LRD) have been studied in several works. In [18] can be fond the first evidence that network traffic has self-similar characteristics. In [23], several TCP statistics, such as session and connection arrivals, were analyzed and self similarity was found in many traces. In [32], the authors provided several measurements which showed that network traffic exhibits self-similar behavior. Physical features of communication networks were also presented to explain such behavior. In [31], time-series extracted from network traffic were proven to exhibit LRD. Feldmann *et al.* investigated several aspects of user and network behaviors contribute to the scaling regimes in WAN traffic [10].

IV. Identification Methodology

Our work aims at classifying the several interactions that an application creates, which may consist of several sessions with different end-hosts/servers. We believe that the analysis of such interactions as a whole can provide a deeper insight into how the applications behave and can assist in traffic discrimination. To be able to perform such study, the classical definition of a flow, the *5-tuple*, becomes too restrictive since it does not capture all the mentioned interactions. Therefore, we used the definition of *data stream*, which consists of all traffic (in the upload or download directions) of a local IP address and univocally identified by a numeric identifier. This *data stream* numeric identifier is: (i) for unencrypted traffic, a specific TCP/UDP (local or remote) port number and (ii) for encrypted traffic, the Security Parameters Index (SPI) in ESP headers in case of IPsec tunnels or any other specific identifier of IP-level encrypted tunnel technology. Therefore, *data streams* are uniquely identified by a *2-tuple* (IP address, unique identifier). Other important definitions in our work are the *known data streams* which consist of *streams*, as previously defined, analyzed *a priori* to determine its origin application(s). On the other hand, let us define the *unknown data streams* as a traffic *stream* created by an unknown application. Several stochastic processes (and respective statistics) can be extracted from these *data streams*, which, in this work, will be processed by a DWT, as described in Section III, in order to obtain the estimators defined in (2). Since the applications that generated the analyzed traffic might have different network conditions, these estimators were normalized to zero mean:

$$\hat{\mu}_j = \mu_j - \sum_{j=1}^{J} \frac{\mu_j}{J} \quad (3)$$

in which J represents the number of scales considered for analysis. In the following lines we will present some more definitions. For instance, let A represent the number of known applications, M represent the number of *unknown data streams* that we want to classify and N correspond to the number of *known data streams*. Let $p_{i,a}, a = 1, ..., A$ designate the probability that the *unknown stream* i belongs to the Gaussian distribution inferred from the *known streams* of the application a. Let $E_{a,j} = \{e^i_{a,j}, i = 1, \ldots, N\}$ and $U_j = \{u^i_j, i = 1, \ldots, M\}$ represent the normalized estimators, as defined in (3), for the first order moment of the wavelet coefficients of a stochastic process, respectively, extracted from a *known data stream* i of the application a at the scale j and extracted from a *unknown data stream* i at the scale j. The proposed methodology assumes that $E^i_{a,j}$ and U_j, for all j and a, follows a Gaussian distribution. Therefore, let

$$P_{i,a,j} = \int_{u^i_j - \Delta}^{u^i_j + \Delta} \frac{1}{\sqrt{2\pi\sigma^2_{a,j}}} e^{\left(\frac{-(u-\overline{e}_{a,j})^2}{2\sigma^2_{a,j}}\right)} du \quad (4)$$

represent the probability that the estimator of the *unknown stream* i, of the scale j, is within a neighborhood of width 2Δ, centered on itself originated by a distribution whose parameters, $\overline{e}_{a,j}$ and $\sigma^2_{a,j}$, are empirically inferred from the *known data streams* of an application a:

$$\overline{e}_{a,j} = \frac{1}{N} \Sigma_{i=1}^{N} e^i_{a,j} \quad (5)$$

$$\sigma^2_{a,j} = \frac{1}{N-1} \Sigma_{i=1}^{N} (e^i_{a,j} - \overline{e}_{a,j})^2 \quad (6)$$

The probability $P_{i,a,j}$ is then computed for all *unknown streams* and for all distributions inferred from the *known streams* studied applications, for each scale of analysis.

Subsequently, it is possible to compute $P_{i,a}$ as:

$$P_{i,a} = \prod_{j=1}^{J} P_{i,a,j}, a = 1, ..., A; i = 1, ..., M \quad (7)$$

Finally, an *unknown data stream* i, $i = 1, \ldots, M$, is associated with application $\alpha, \alpha = 1, \ldots, A$, such that

$$\exists \alpha, P_{i,\alpha} = \max_a [P_{i,a}]. \quad (8)$$

V. Results

In this Section we present the obtained results from several traffic *data streams* extracted from: (i) licit TCP and UDP traffic traces passively collected at the University of Aveiro network on September 15, 2008 and (ii) illicit traces experimentally generated in laboratory simulating some of the most relevant *botnet* uses. The licit applications *data streams* extracted (and classified *a priori*) from the traffic collected were file-sharing (BitTorrent), video streaming and HTTP (browsing). Figures 1 to 3 present the variation of the number of bytes in the upload and download directions for the mentioned applications. The illicit traffic was experimentally generated in our lab in an attempt to simulate some of the most relevant reconnaissance attacks. The NMAP [1] flows were generated using a discrete scan profile in order to replicate a typical *botnet* port scan that tries to evade IDS detection and

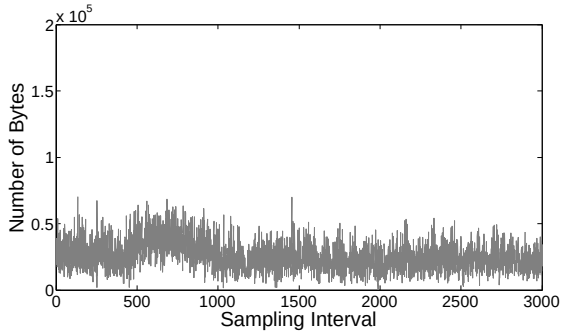

Figure 1. Number of bytes for a Torrent flow.

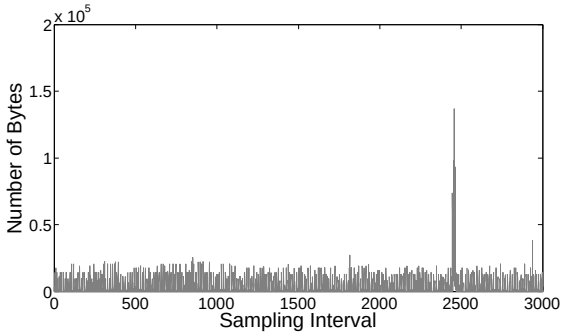

Figure 2. Number of bytes for a Streaming flow.

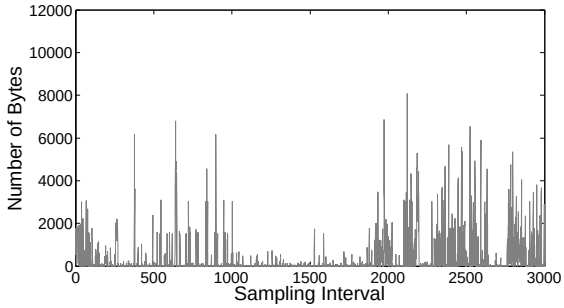

Figure 3. Number of bytes for an HTTP flow.

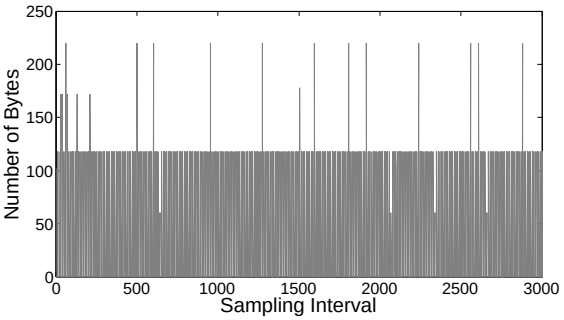

Figure 4. Number of bytes for a NMap flow.

scan hosts and networks, bypassing their firewalls and proxies. Therefore, we performed a sequential port scan with one second of interval between (SYN) probes and a waiting time of 15 seconds before start scanning a new machine. The Snapshot flows were generated by emulating the capture of a fixed size small image (335x180 pixels, 120KBytes) of the user's desktop around the cursor every time the user performed a click. We assumed that the user was browsing the Internet and performed a click with an exponentially distributed interval with average equal to 120 seconds [33]. The flows of these applications are presented in Figs. 4 and 5, respectively.

In this case, the values analyzed were the overall number of transmitted bytes, independently of direction. The extracted *data streams* were 5 and 15 minutes long and were divided in *known* and *unknown streams*, however, the real classification of all *streams* was kept for validating the classification results of the *unknown streams*. Now let us present the values of the several variables defined in Section IV. The number of application considered was 5 ($A = 5$). The number of *known streams*, N, used for inferring the parameters of the Gaussian distributions was 30 and the number of *unknown streams*, M, was 80, for each application. The value of the interval Δ used was 0.1.

The *known* and *unknown streams* were analyzed via a DWT in order to obtain the estimators for the first order moment of the wavelet coefficients. The first mentioned values were then used to validate the assumption that the estimators for the first order moment of the wavelet coefficients, for each application and scale, follow a Gaussian distribution. The test used was the Lilliefors goodness-of-fit test which verifies the null hypothesis that the sample in a vector comes from a distribution in the Gaussian family, against the alternative that it does not [19]. All the tests did not reject the null hypothesis, that is, all the estimators can be approximated by a Gaussian distribution.

The classification results were computed by comparing the classification achieved with the proposed methodology with the real application. In the first part of our results, we considered the 5 minutes long *data streams*. We only used the first 5 scales since at higher scales the estimators of all applications tend to converge. Figures 6 and 7 show box plots with 25%, 50%, 75% and 95% quantiles, for the estimators of the first order moment of the wavelet coefficients of the 5 minutes and 15 minutes *data streams*, respectively. We can observe that the distributions of the estimators of the HTTP and Snapshot *streams* almost overlap in all scales. This suggests that some HTTP and Snapshot *streams* might be misclassified. However, for the 15 minutes *data streams* (Figure 7) the Snapshot traffic estimators are now more concentrated around the mean, which suggests that the accuracy will be higher. For the remaining estimators' distributions we can observe that, at least in one scale, they are very separated and therefore they will not be misclassified.

The numerical results obtained, for the 5 minutes traffic traces, are presented in Table I and it is possible to observe that these are relatively accurate for all applications. With the exception of HTTP 5 minutes *data streams*, the obtained percentage of correctly identified *data streams* is between 73% and 100%. For HTTP traffic, the correct classification percentage is lower, as some of these *data streams* were misclassified as Snapshot, which is in accordance with the previous analysis. These result can be explained by the fact

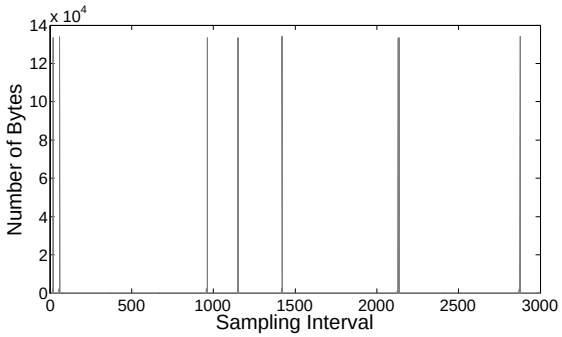

Figure 5. Number of bytes for a Snapshot flow.

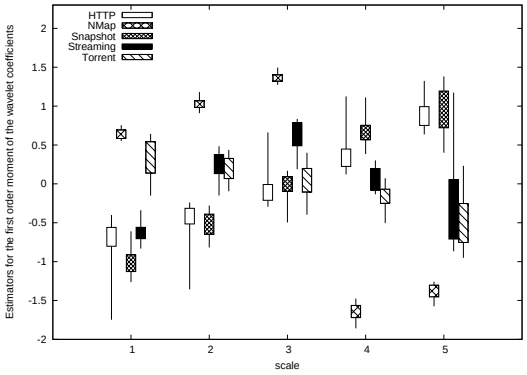

Figure 6. Distributions for 5 minutes traces.

that HTTP *data streams* multiscale estimators have an higher variance, resulting from the various and heterogeneous user behaviors, making this distribution partially overlap the snapshot estimators distribution (which has a much lower variance) in all scales. Moreover, several protocols, such as file sharing and video streaming, run on top of HTTP communications which justifies the large variance the estimators of these *streams* present and some classification mistakes. The classification results for the 15 minutes *data streams* are presented in Table II and we can observe that the accuracy of the results for all applications is higher. This can be explained by the fact that traces are longer, contain more information and more differentiating characteristics. This allows a deeper decomposition of each signal and therefore, a better analysis of their unique behaviors and leads to better classification results.

Table I
RESULTS FOR 5 MINUTES TRACES USING 5 SCALES

Data Streams	Classified as *NMap*	*Snapshot*	*HTTP*	*Streaming*	*Torrent*
NMap	**100%**	0%	0%	0%	0%
Snapshot	0%	**72.7%**	22.7%	3.1%	1.5%
HTTP	0%	29.4%	**64.7%**	2.9%	2.9%
Streaming	0%	0%	3.6%	**96.4%**	0%
Torrent	0%	3.1%	1.6%	1.5%	**93.8%**

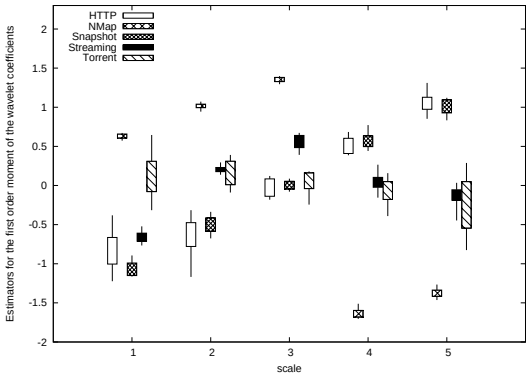

Figure 7. Distributions for 15 minutes traces.

Table II
RESULTS FOR 15 MINUTES TRACES USING 5 SCALES

Data Streams	Classified as *NMap*	*Snapshot*	*HTTP*	*Streaming*	*Torrent*
NMap	**100%**	0%	0%	0%	0%
Snapshot	0%	**95.2%**	4.8%	0%	0%
HTTP	0%	15.4%	**76.9%**	0%	7.7%
Streaming	0%	0%	0%	**100%**	0%
Torrent	0%	0%	0%	0%	**100%**

VI. CONCLUSIONS

The last years have witnessed the appearance of several new protocols and services, a huge increase on the number and variety of security vulnerabilities and attacks that are carried out over the Internet and the growth of the privacy and confidentiality concerns of Internet users. Thus, new identification methodologies that can be accurate when applied to different types of traffic and be able to operate in cyphered traffic scenarios are needed. This paper proposed an identification methodology that relies on a statistical multiscale analysis of the traffic flows, differentiating them based on the probability that their characteristic multiscale behavior estimators belong to Gaussian probability distributions whose parameters are inferred from traffic flows of real applications. The results obtained show that the proposed methodology is able to accurately classify licit traffic and also identify some of the most common Internet security attacks. Besides, the approach can also avoid some of the most important drawbacks presented by existing identification methodologies, namely their inability to work under strict confidentiality restriction scenarios. Finally, the definition of *data stream* also proved to be adequate for discriminating between several IP applications, constituting an important step towards a complete understanding of their behaviors.

ACKNOWLEDGEMENTS

This research was supported in part by Fundao para a Cincia e a Tecnologia, grant SFRH/BD/33256/2007.

REFERENCES

[1] Nmap: Free security scanner for network exploration and security audits. http://nmap.org/, March 2009.

[2] Snort :: Home page. http://www.snort.org/, May 2010.

[3] P. Abry, P. Flandrin, M. Taqqu, and D. Veitch. Wavelets for the analysis, estimation, and synthesis of scaling data. In K. Park and W. Willinger, editors, *Self-Similar Network Traffic and Performance Evaluation*, pages 39–88. Wiley, 2000.

[4] P. Barford and V. Yegneswaran. An inside look at botnets. *Springer Verlag*, 2006.

[5] K. Chiang and L. Lloyd. A case study of the rustock rootkit and spam bot. In *HotBots'07: Proceedings of the first conference on First Workshop on Hot Topics in Understanding Botnets*, pages 10–10, Berkeley, CA, USA, 2007. USENIX Association.

[6] M. Collins, T. Shimeall, S. Faber, J. Janies, R. Weaver, M. Shon, and J. Kadane. Using uncleanliness to predict future botnet addresses. In *ACM/USENIX Internet Measurement Conference IMC'07*, 2007.

[7] E. Cooke, F. Jahanian, and D. Mcpherson. The zombie roundup: Understanding, detecting, and disrupting botnets. pages 39–44, June 2005.

[8] E. Cooke, F. Jahanian, and D. McPherson. The zombie roundup: Understanding, detecting, and disrupting botnets. In *USENIX SRUTI'05*, pages 39–44, 2005.

[9] D. Dagon, C. Zou, and W. Lee. Modeling botnet propagation using timezones. In *13th Annual Network and Distributed System Security Symposium NDSS'06*, January 2006.

[10] A. Feldmann, A. Gilbert, P. Huang, and W. Willinger. Dynamics of IP traffic: A study of the role of variability and the impact of control. In *SIGCOMM*, pages 301–313, 1999.

[11] A. Feldmann, A. C. Gilbert, and W. Willinger. Data networks as cascades: investigating the multifractal nature of internet wan traffic. In *SIGCOMM '98: ACM SIGCOMM '98 conference on Applications, technologies, architectures, and protocols for computer communication*, pages 42–55, New York, NY, USA, 1998.

[12] G. Gu, P. Porras, V. Yegneswaran, M. Fong, and W. Lee. Bothunter: Detecting malware infection through ids-driven dialog correlation. In *16th USENIX Security Symposium*, 2007.

[13] Y. Hu, D.-M. Chiu, and J. Lui. Application identification based on network behavioral profiles. *Quality of Service, 2008. IWQoS 2008. 16th International Workshop on*, pages 219–228, June 2008.

[14] T. Karagiannis, A. Broido, N. Brownlee, K. Claffy, and M. Faloutsos. Is p2p dying or just hiding? [p2p traffic measurement]. *Global Telecommunications Conference, 2004. GLOBECOM '04. IEEE*, 3:1532–1538 Vol.3, Nov.-3 Dec. 2004.

[15] T. Karagiannis, K. Papagiannaki, and M. Faloutsos. Blinc: multilevel traffic classification in the dark. In *SIGCOMM '05: 2005 conference on Applications, technologies, architectures, and protocols for computer communications*, pages 229–240, New York, NY, USA, 2005. ACM.

[16] A. Karasaridis, B. Rexroad, and D. Hoeflin. Wide-scale botnet detection and characterization. In *HotBots'07: First Workshop on Hot Topics in Understanding Botnets*, Berkeley, CA, USA, 2007. USENIX Association.

[17] O. Kolesnikov, D. Dagon, and W. Lee. Advanced polymorphic worms: Evading ids by blending in with normal traffic. Technical report, 2004.

[18] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the self-similar nature of ethernet traffic (extended version). *IEEE/ACM Transactions on Networking*, 2(1):1–15, 1994.

[19] H. Lilliefors. On the kolmogorov-smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 1967.

[20] C. Livadas, R. Walsh, D. Lapsley, and W. Strayer. Using machine learning techniques to identify botnet traffic. In *2nd IEEE LCN Workshop on Network Security*, 2006.

[21] A. Madhukar and C. Williamson. A longitudinal study of p2p traffic classification. *Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, 2006. MASCOTS 2006. 14th IEEE International Symposium on*, pages 179–188, Sept. 2006.

[22] A. H. Nicholas Ianelli. Botnets as a Vehicle for Online Crime. *The International Journal of Forensic Computer secience*, 2(1), 2007.

[23] V. Paxson and S. Floyd. Wide-area traffic: The failure of Poisson modeling. *IEEE/ACM Transactions on Networking*, 3(3):226–244, June 1995.

[24] M. Rajab, J. Zarfoss, F. Monrose, and A. Terzis. A multi-faceted approach to understanding the botnet phenomenon. In *ACM SIGCOMM/USENIX Internet Measurement Conference IMC'06*, October 2006.

[25] A. Ramachandran and N. Feamster. Understanding the network-level behavior of spammers. *SIGCOMM Comput. Commun. Rev.*, 36(4):291–302, 2006.

[26] A. Ramachandran, N. Feamster, and D. Dagon. Revealing botnet membership using dnsbl counterintelligence. In *USENIX SRUTI*, 2006.

[27] M. Reiter and T. F. Yen. Traffic aggregation for malware detection. In *Fifth GI International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*, 2008.

[28] E. Rocha, P. Salvador, and A. Nogueira. Detection of illicit traffic based on multiscale analysis. In *Software, Telecommunications Computer Networks, 2009. SoftCOM 2009. 17th International Conference on*, pages 286 –291, September 2009.

[29] S. E. Security. Symantec Global Internet Security Threat Report: Trends for 2008. Technical report, Symantec, April 2009.

[30] S. Sen, O. Spatscheck, and D. Wang. Accurate, scalable in-network identification of p2p traffic using application signatures. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 512–521, New York, NY, USA, 2004. ACM.

[31] M. Taqqu, V. Teverovsky, and W. Willinger. Is network traffic self-similar or multifractal? 5:63–73, 1997.

[32] W. Willinger, V. Paxson, and M. Taqqu. *Self-similarity and Heavy Tails: Structural Modeling of Network Traffic*. A Practical Guide to Heavy Tails: Statistical Techniques and Applications. Birkhauser, 1998.

[33] I. Zukerman, D. W. Albrecht, and A. E. Nicholson. Predicting users' requests on the www. In *Seventh International Conference on User Modeling*, pages 275–284, 1999.

 ISBN: 978-1-61208-103-8