

# Iterative packet-based approach for early application identification

Mohamad JABER, Chadi BARAKAT

Project-Team Planète, INRIA Sophia-Antipolis, France

{mohamad.jaber, Chadi.Barakat} @sophia.inria.fr

## Abstract

One of the most important challenges for network administrators is the identification of applications in Internet traffic. In practice, there is a need to classify Internet traffic for many purposes in network security, traffic engineering and monitoring. The classical method based on standard port numbers is less and less efficient given the large number of applications using non standard ports. Methods based on deep packet inspection are known to be slow on high speed networks and unable to identify encrypted traffic. In this work, we study the feasibility of classifying Internet traffic based on statistical properties as the packet size. We come up with an iterative probabilistic method able to identify Internet traffic quickly and accurately by only using the size of the first N packets. Our main observation is that the accuracy of our method increases while increasing the number of observed packets.

**Keywords:** Application identification, Traffic measurements

## I. INTRODUCTION

Network traffic is composed of many applications, including Internet-like applications ((Web, mail, online games, P2P...) and other proprietary ones. Faced by this diversity, Internet Service Providers (ISPs) are more and more interested in identifying the different applications sources of the traffic. This will allow a precise application reporting, relevant indicators on QoS and SLA (depending on the applications), and consequently a good control and dimensioning of the network. This will also facilitate the troubleshooting (detect and locate a performance problem in the network), and will help to filter unknown and possibly harmful applications, to prioritise some major applications, and to evaluate if users and applications encounter the right and satisfactory quality of service and if the service level agreements are respected. The recognition of applications should help to strengthen our knowledge on applications, in order to take the right decision for the control of the quality and costs of the network.

The recognition of these applications on IP traces becomes increasingly complex. Historically the recognition starting

from the static, preset and standard port numbers met the need well. But some applications now use dynamic port numbers; this is typically the case of telephony over IP. Other applications hide themselves using standard ports stolen from other applications. These ports are usually given by the end host and thus they can be easily changed.

Current techniques of "Deep Packet Inspection" (DPI) [1]-[6] make it possible to go further in the identification of the applications but they require a complete and costly exploration of the payload of the packets. This induces an important load and requires updates with the appearance of new applications. Furthermore, when packets are encrypted, the recognition is not possible.

The Statistical techniques [8]-[18] seem to be today an interesting alternative. They allow to recognize and to classify the applications according to their statistical signatures. These signatures can be volumes (number of bytes) per connection, connection durations, rates, inter-packets delays, packet sizes, and direction. But the majority of these methods can't identify flows early and they require to reach the end of the flow before taking the decision which could be too late for some applications related to network administration.

In [18], McGregor et al. show the utility of using clustering algorithms for the identification of the traffic while using unsupervised machine learning called auto class and the following statistical criteria: packet size, inter-arrival time, byte count and connection duration. In [16], Moore et al. use a supervised machine learning called Naïve Bayes to classify the TCP traffic, and they try to find the best set of statistical criteria. In [10], Bernaille et al. use three clustering algorithms (K-Means, Gaussian mixture model and the Spectral clustering) while using the size and the direction of the first four packets to assign flows to clusters, but they use ultimately the port number as a factor to differentiate between flows falling in the same cluster.

In [8] Erman et al. make a comparison between three unsupervised machine learning techniques (K-Means, DBSCAN and Auto-class), but they classify connections after their end. Our method is based on the same approach with the difference that we precede iteratively packet per packet starting from the beginning of a new flow, and we take the

decision when a certain confidence level is reached.

## II. METHOD DESCRIPTION

Our method is a statistical method allowing iterative early classification of internet traffic. The metrics we track are the size and the direction of the first N packets of a flow and a predefined controlled confidence level in the port number. This latter parameter is one of the novelties brought by our method compared to the literature. For each new packet from a flow, we update our confidence about the different decisions we can take. Then value of N at which to stop is a function of the level of precision and confidence we target.

Our method consists of three phases the model building phase where we build several clusters and where each cluster is labelled by a specific application, the classification phase where each flow is affected in a class among the classes of the training data set according to the similarity based in the Euclidian distance and the phase of computing probability and Labeling flows with applications. We calculate the probability that a flow F belongs to an application I knowing the result of classification as follow:

$$Pr(I/Result) = \frac{Pr(I) * \prod_{k=1}^n Pr(i(k)/I)}{\sum_{I=1}^{aPP} Pr(I) * \prod_{k=1}^n Pr(i(k)/I)}$$

$Pr(i(k)/I)$  is the proportion of the application I in the class i during the test K,  $Pr(I)$  is the proportion of the application I generally across the sample on which we work, n is the number of test and app is the number of applications.

We used a known clustering algorithm called KMeans that classifies connections in clusters according to the similarity. Let us note that for KMeans we specifies at the beginning the number of clusters which we wants to obtain..

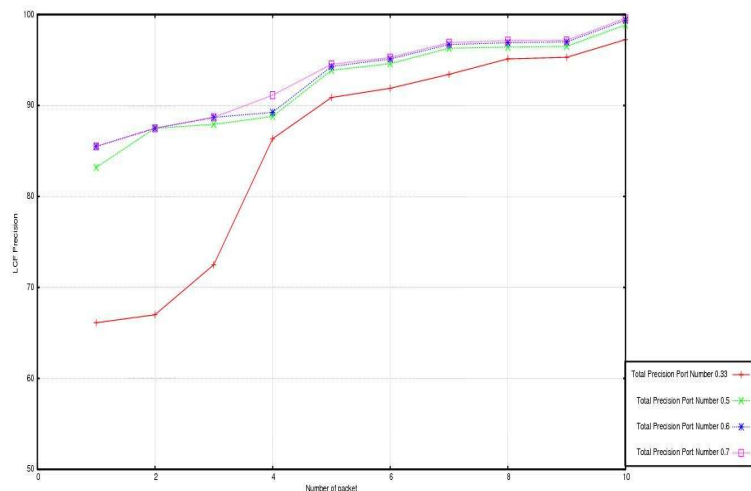
In the classification phase and for each new connection we calculate the Euclidean distance between this connection and the center of each cluster and we assign the connection to the nearest one.

We made a learning phase and a classification phase depending on the size and the direction of each package alone among the first packets N, and after we calculated the probability of belonging to each application following the equation, and we assign the flow to the most likely application if the level of confidence is reached, if not we add a new packet iteratively.

## III. RESULTS

We compute Precision for our classification by dividing the total number of connections well classified by the total number of the connections.

We validate our method in a real internet trace collected from INRIA Sophia Antipolis and we found very good result, and from the 10th we reached 97% without any influence by the port number (which corresponds to 0.33 in the graph).



## IV. CONCLUSION AND FUTURE WORK

In conclusion we developed a new and simple method that allows identifying TCP and UDP traffic, with little meadows in real times.

First results seem very encouraging, and to further validate our method and prove its efficiency, we now plan to test it on other networks (ADSL network) and on several Data sets that contain different mix of applications.

## References

- [1] T. Karagiannis, A. Broido, M. Faloutsos, and K. claffy, "Transport Layer Identification of P2P Traffic," in *IMC'04*, Taormina, Italy, October 25-27, 2004.
- [2] P. Haffner, S. Sen, O. Spatscheck, and D. Wang, "ACAS: Automated Construction of Application Signatures," in *SIGCOMM'05 Workshops*, Philadelphia, USA, August 22-26, 2005.
- [3] A. W. Moore and K. Papagiannaki, "Toward the Accurate Identification of Network Applications", In Proceedings of the 6th Passive and Active Measurement Workshop (PAM 2005), pages 41–54, October 2005
- [4] D. Antoniadis, M. Polychronakis, S. Antonatos, E. Markatos, S. Ubik & A. Øslebø, "Appmon: An Application for Accurate per Application Network Traffic Characterization" - Proceedings of IST BroadBand Europe 2006, December 2006, Geneva, Switzerland
- [5] S. Sen, O. Spatscheck, and D. Wang, "Accurate, Scalable In-Network Identification of P2P Traffic Using Application Signatures", in WWW 2004 Conference New York, USA, May 17-22, 2004.
- [6] M. Perényi, T. Dinh Dang, A. Gefferth, S. Molnár, "Identification and Analysis of Peer-to-Peer Traffic" in JOURNAL OF COMMUNICATIONS, VOL. 1, NO. 7, November/December 2006
- [7] T. Karagiannis, K. Papagiannaki, and M. Faloutsos, "BLINK: Multilevel Traffic Classification in the Dark," in *SIGCOMM'05*, Philadelphia, USA, August 21-26, 2005.
- [8] J. Erman, M. Arlitt, A. Mahanti, "Traffic Classification Using Clustering Algorithms", Proceedings of the 2006 SIGCOMM workshop on Mining network data, Pisa (Italy), pages 281-286, September 2006
- [9] J. Erman, A. Mahanti, and M. Arlitt, "Internet Traffic Identification using Machine", Proc. of 49th IEEE Global Telecommunications Conference (GLOBECOM), San Francisco, USA, November 2006

- [10] L. Bernaille, R. Teixeira, and K. Salamatian, "Early Application Identification, In The 2nd ADETTI/ISCTE CoNEXT Conference, Lisboa, Portugal, December 2006.
- [11] L. Bernaille, R. Teixeira, I. Akodkenou, A. Soule† and Kave Salamatian "Traffic Classification On The Fly", in Computer Communication Review (Editorial), April 2006.
- [12] N. Williams, S. Zander, G. Armitage "A preliminary performance comparison of five Machine Learning for practical IP Flow classification" in ACM SIGCOMM computer communication review October 2006
- [13] M. Crotti, M. Dusi, F. Gringoli and L. Salgarelli, « Traffic Classification through simple statistical Fingerprinting", ACM-Sigcomm Computer Communication Review, Volume 37, Issue 1, Pages 5-16, January 2007
- [14] S. Zander, T. Nguyen, G. Armitage, "Self-learning IP Traffic Classification based on statistical Flow Characteristics" Proc. Passive and Active Measurement workshop (PAM 2005), Boston, MA (USA), March/April 2005
- [15] D. Zuev and A. Moore "Traffic Classification using a statistical approach" in Passive & Active Measurement Workshop, Boston, U.S.A, April 2005
- [16] A. Moore and D. Zuev, "Internet Traffic Classification Using Bayesian Analysis Techniques" in *Sigmetrics 2005*
- [17] S. Zander, T. Nguyen and G. Armitage, "Automated traffic classification and application identification using Machine Learning" in IEEE Conference 2005
- [18] A. McGregor, M. Hall, P. Lorier, and J. Brunskill "Flow clustering using Machine Learning Techniques" in Passive & Active Measurement Workshop, France, April 2004.
- [19] M. Eisen, P. Spellman, P. Brown, and D. Botstein. "Cluster Analysis and Display of Genome-wide Expression Patterns." *Genetics*, 95(1):14863.15868, 1998.
- [20] I. Witten and E. Frank. (2005) "*Data Mining: Practical Machine Learning Tools and Techniques*." Morgan Kaufmann, San Francisco, 2nd edition, 2005.
- [21] P. Cheeseman and J. Strutz. "Bayesian Classification (AutoClass): Theory and Results." *In Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press, USA*, 1996.
- [22] A. Moore, and D. Zuev, "Discriminators for use in flow-based classification." Technical Report IRC-TR-04-028, Intel Research, Cambridge (2004)
- [23] IANA. Internet Assigned Numbers Authority (IANA), <http://www.iana.org/assignments/port-numbers>.
- [24] TCPdump, <http://www.tcpdump.org/>