Emilie: Enhance the Power of Traffic Identification

Yiyang Shao[*†], Baohua Yang[*†], Jingjie Jiang[*], Yibo Xue[†‡¶] and Jun Li[†‡]
[*]Department of Automation, Tsinghua University, Beijing, China
[†]Research Institute of Information Technology, Tsinghua University, Beijing, China
[‡]Tsinghua National Lab for Information Science and Technology, Beijing, China
{shaoyy11, ybh07, jjx08}@mails.tsinghua.edu.cn, {yiboxue, junl}@tsinghua.edu.cn

***Abstract*—Network traffic identification has become more and more important in recent years. However, as the Internet backbone bandwidth continuously grows, traditional flow-based traffic identification methods gradually become impractical. In order to improve the performance of traffic identification, this paper proposes an ingenious and practical flow dispatching mechanism named Emilie, which intelligently predicts the elephant flows using only the first three packets of each flow. By discriminating mouse flows against elephant flows, methods with various complexity are utilized to identify the application-level protocol type of elephant and mouse flows separately. Emilie utilizes Machine Learning techniques to achieve high accuracy as well as keep fast speed in predicting elephant flows. Experimental results on real network traffic traces illustrate that around 88% precision, 85% recall and over 85% accuracy are gained on average, which is much better than existing solutions. To the best of our knowledge, this is the first practical and efficient work that supports inline elephant flow prediction. Flow dispatching based on Emilie empowers traffic identification systems to achieve both high accuracy and fast speed.**

***Index Terms*—Elephant Flows Prediction; Traffic Identification; Flow Dispatch**

I. Introduction

The Internet has been dramatically changing ever since it appears, and lots of challenges are emerging in the network traffic management [1]. These issues have motivated many novel approaches, among which a fundamental one is traffic identification. Typically, the aim of traffic identification is to classify the traffic into its corresponding protocols. Traffic identification reveals protocols, and thus provides important supports to traffic management, *e.g.,* measurement and control. Internet Service Providers (ISP) often rely on traffic identification to analyze and optimize their networks [2].

Many traffic identification works focus on the flow level [3]–[6], for the reason that the aggregation from packets to flows effectively reduces the processing complexity while guaranteeing enough fine-granularity. These works can be mainly categorized into three classes based on their motivations.

The header-based method is the most fundamental method. In this method, the identification is achieved based on features in packet headers, among which a typical one is well-known destination port numbers given by IANA [7]. However, since many applications take un-reserved port numbers or tunneled in common ports, this method cannot guarantee the accuracy.

Relying on the Deep Inspection (DI) techniques, payload-based method becomes today's most accurate solution to traffic identification. However, its disadvantage is that payload scanning results in a slow speed compared with header-based method. What's more, encryption of traffic will cause identification failures. At last, there are restrictions to third party payload inspection under the privacy regulations [8].

An emerging way is statistics-based method, which employs Machine Learning techniques [5]. This method identifies different applications utilizing flow statistical information, such as packet inter-arrival time and packet size [6]. However, statistical information may vary in different network environments, and thus a classifier suitable for one specific environment may become invalid in others. Besides, dynamic protocols, especially some P2P protocols, may deliberately perform similar statistics to cause misjudgements.

In light of the above analysis, all existing methods still face various problems in practical deployment. Recently, it has been suggested to combine different methods together to realize practical performance [6], but the remaining question is how to dispatch the flows to different types of approaches. It is well known that very few large flows, namely the *elephant flows* (elephants for short), are responsible for a high percentage of the traffic volume, and on the contrary, a very high percentage of the flows, namely the *mouse flows* (mice for short), are responsible for a few percentage of the traffic volume. This is called the "elephants and mice phenomenon" [9]–[12]. Based on this observation, it is promising to take the sophisticated methods (*e.g.,* payload-based) to identify those elephants and achieve accurate results of large amount of traffic volume. In this way, the performance of traffic identification can be further improved by the prediction of elephants and mice for subsequent detection. However, most existing methods for elephant detection are designed for off-line classification (See Section II), and it is difficult to employ them for online traffic identification, which requires an early identification.

In this paper, we propose Emilie, an *Elephants and MIce fLow dIspatchEr*, which achieves an early prediction at the beginning of each flows. With the lengths of the first three packets of each flow as features, Emilie utilizes Machine Learning techniques to detect elephants, and thus achieves high accuracy while keeping fast speed. In this way, traffic identification system based on Emilie can organically combine fine-granular and coarse-granular devices together. To the best of our knowledge, this is the first practical and efficient work that identifies elephants early enough, so that the flows can be dispatched effectively for subsequent processing.

[¶]Corresponding author. E-mail: yiboxue@tsinghua.edu.cn

978-1-4799-2358-8/14/$31.00 ©2014 IEEE

Main contributions of this paper include:

- *Early identification*. Emilie achieves early identification of elephants with only the lengths of first three packets in each flow, which is crucial for online deployment.
- *Fast speed*. Our approach requires no packet payload detection, which guarantees very fast processing speed.
- *Remarkable efficiency*. Integrated with Emilie, existing traffic identification approaches gain significant speedup.

The rest of this paper is organized as follows. Section II introduces the related work. Section III describes the design of Emilie. Evaluation and analysis of experimental results are shown in section IV. Some issues about Emilie are discussed in section V. Finally in section VI, we concludes the paper.

II. Related Work

Limited by storage and computation resource, it is impossible to collect and monitor all network packets. Many approaches aim to aggregate information from different angles, among which, a fundamentally effective approach is the concept of elephants and mice. According to statistics, the top 9% of flows between Autonomous Systems (ASes) contribute to 90.7% of traffic in terms of bytes [9]. Another observation in campus network shows a similar result that the top 5% of flows contribute to over 83% of traffic volume [13]. In many areas, especially traffic engineering, what's really important is flows' volume but not quantity. Therefore, the classification of elephants and mice plays an important role in these areas.

Traditionally, main solutions to identify elephants include counting-based, hash-based and sampling-based methods. In counting-based method, a limited number of counters are used to find frequent items in a data stream [10]. In hash-based method, one or two dimension counters are used to construct a hash table to estimate the frequencies of different items [11]. And finally in sampling-based method, the frequency of items are estimated by periodically sampling in data stream [12]. Though these three kinds of methods can achieve high accuracy with low complexity, they identify elephant flows only after vast traffic volume passed through, and cannot provide early prediction. In contrast, Emilie is designed with fundamentally different intention. To provide an early classification, we propose Emilie for intelligent prediction of elephants utilizing Machine Learning method.

III. System Design

As a system, Emilie is designed to interact with a few functional components, and at its core is the SVM classifier to make it efficient and practical.

A. Framework

As mentioned above, by utilizing the length feature of each flow, Emilie constructs a classifier to predict whether or not a flow is an elephant. Classification of a flow involves a number of steps. First, features are defined as the sizes of the first three packets of each flow. Then training dataset is required to associate sets of features with known classes (elephant or mouse). Finally we apply the classifier to predict the classes of unlabeled flows based on their length features. According to the series of procedures above, Emilie can be divided into four components: Metering, Training, Classifying and Dispatching. Fig. 1 shows the framework and a brief description of each component is shown below.

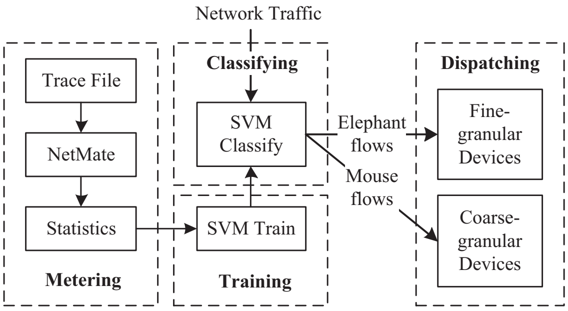

Fig. 1. Framework of Emilie

1) Metering: The main function of the Metering component is to generate training dataset. By adding a flow management module to the open-source metering software NetMate, the Metering component extracts flow statistics including length feature and total size of each flow in the off-line trace files. TCP and UDP packets are sequentially collected from the trace files to preserve real life cases of packet disorder.

2) Training: This component constructs the SVM classifier based on the flow statistics obtained by the Metering component. A sample flow is labeled as elephant when its total size is above a flow size threshold. Same number of elephants and mice labeled samples are randomly chosen to generate the training dataset, which is then employed to construct the SVM classifier. Noticeably, the number of mice is much larger than that of elephants in the trace files. So if we simply use all samples in the trace files, the decision boundary will be biased towards the mice class and thus cause accuracy degradation.

3) Classifying: The main function of the Classifying component is to distinguish elephants and mice in online traffic according to the classifier. To measure the accuracy of the classifier, testing dataset are generated with the labeled flow samples other than those in the training dataset.

4) Dispatching: Leveraging the flow classification result given by SVM classifier, this component dispatches different classes of flows to fine-granular and coarse-granular devices to further identify the traffic. Fine-granular devices are utilized to process the elephants while coarse-granular devices are prepared for mice, and thus a slow but accurate result will be contributed to a few flows with large amount of traffic volume.

B. SVM Construction

SVM classifier plays the most important role in the Emilie system. There are two key processes during the construction of SVM classifier, which significantly impacts the overall identification accuracy.

1) Feature Selection: Length vector of the first several packets is selected as the classification feature with the following consideration. First, our design aims to achieve an early

TABLE I
Symbol Statistics of Internet Traffic

	Volume	**Amount**	**Mean Flow Size**
Total	V	N	m
Elephants	$\alpha \cdot V$	$\beta \cdot N$	$\frac{\alpha}{\beta} \cdot m$
Mice	$(1-\alpha) \cdot V$	$(1-\beta) \cdot N$	$\frac{1-\alpha}{1-\beta} \cdot m$

TABLE II
Symbol Statistics of Traffic Identification Devices

	Bps & fps	**Accuracy**
Emilie based device	B_e & F_e	η_e
Fine-granular device	B_1 & F_1	η_1
Coarse-granular device	B_2 & F_2	η_2

classification, so the features must be gained at the beginning of a flow. Thus, features like flow size or duration time are not suitable. Second, the features should not be network environment dependent. Features that are heavily influenced by the variation of network condition, for example the packets inter-arrival time, will lead to a degraded efficiency in the case of congestion. The last and the most important reason is that the features must have essential relationship with the classes. We will discuss this further in the discussion part of evaluation section.

2) Threshold Selection: The determination of traffic volume threshold, which is the fundamental baseline to distinguish elephants from mice, depends on network traffic characteristic and deep inspection throughput. Generally, if considering minimizing the need for deep inspection processing power and only selecting the threshold based on normal traffic characteristic, the value of threshold is often associated with the mean value of flow sizes. In this paper, we introduce the concept of Cumulative Distribution Figure (CDF) of traffic volume to make a better choice of the threshold. CDF describes the relationship between cumulative traffic volume and flow numbers, and will be used for the calculation of threshold in the evaluation part.

C. Speedup Ratio

Leaving the accuracy of Emilie to be fully evaluated in the next section, the speedup ratio of this traffic identification system is compared to original fine-granular or coarse-granular device for the evaluation of Emilies efficiency.

Considering a real-time traffic identification powered by Emilie, we assume that in a certain period, the statistic features of Internet traffic consist of traffic volume, flow amount and mean flow size, which are shown in TABLE I. By comparing the performances of traffic identification approach based on Emilie with traditional fine-granular and coarse-granular ones, the speedup ratio can be calculated. Statistic features including throughput and accuracy of different kind of devices are shown in TABLE II. We use both bytes per second (Bps) and flows per second (fps) as measure units to evaluate the throughput from different angles. And the accuracy represents byte accuracy. Noticeably, under the assumption that the Internet traffic has a steady mean flow size, namely m in TABLE I, we deduce the following relationships: $B_e = F_e \cdot m$, $B_1 = F_1 \cdot m$ and $B_2 = F_2 \cdot m$.

In our design, the elephants are identified by the fine-granular device and the mice are processed with the coarse-granular device. According to the parameters given by TABLE II, the processing time of fine-granular device and coarse-granular device can be calculated as follows:

$$t_1 = \frac{\beta \cdot N}{F_1} \qquad t_2 = \frac{(1-\beta) \cdot N}{F_2}$$

Generally speaking, the system performance depends on its fine-granular device. To make the fine-granular device work at full load, we have $t_1 \geq t_2$ (in optimal condition we have $t_1 = t_2$), and thus the relationship between fine-granular and coarse-granular device is as follows:

$$\beta \cdot F_2 \geq (1-\beta) \cdot F_1 \qquad \beta \cdot B_2 \geq (1-\beta) \cdot B_1$$

The inequations above can be easily satisfied by adjusting the parameters of coarse-granular device. Therefor, the processing time of Emilie is equal to that of the fine-granular device in Emilie, *i.e.*, $t_e = t_1$. Knowing parameters in TABLE I and the processing time, we can calculate Emilie's throughput, accuracy and speedup ratio as follows:

$$F_e = \frac{N}{t_e} = \frac{F_1 \cdot N}{\beta \cdot N} = \frac{F_1}{\beta} \qquad B_e = F_e \cdot m = \frac{F_1 \cdot m}{\beta} = \frac{B_1}{\beta}$$

$$\eta_e = \frac{\eta_1 \cdot \alpha \cdot V + \eta_2 \cdot (1-\alpha) \cdot V}{V} = \eta_1 \cdot \alpha + \eta_2 \cdot (1-\alpha)$$

$$speedup\ ratio = \frac{F_e}{F_1} = \frac{B_e}{B_1} = \frac{1}{\beta}$$

To characterize the speedup ratio and accuracy of Emilie clearer, we assign exact value to the symbols in the equations above. As mentioned above, generally, we consider that the proportion of elephants is 5% in quantity and 90% in volume, which accords with the results in evaluation section. Namely, we have $\alpha = 0.90$ and $\beta = 0.05$. We assume that a fine-granular device can achieve 95% accuracy, and a coarse-granular device's accuracy is 70% [8]. Thus, Emilie can achieve 20 times speedup compared with the process solely using fine-granular device. Meanwhile, the accuracy remains 92.5%, which is almost the same with fine-granular device.

D. Discussion

Obviously, the dispatching of flows will bring about overheads to the traffic identification. We evaluate the overheads of Emilie from two angles: latency and throughput. In terms of latency, the main reason is that we have to wait the arrival of three packets and then begin the classification. However, demonstrated by the experimental results and analytical conclusion, the latency is at microsecond level, which is negligible for subsequent identification. Meanwhile, as shown in the next section, the throughput of Emilie is large enough compared with the throughput of traffic identification devices. Therefore, Emilie is designed with reasonable overhead and the system is practical.

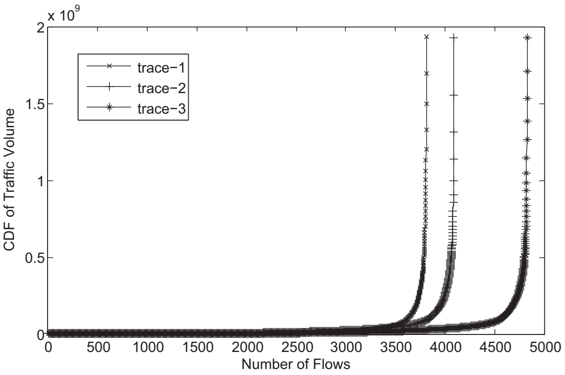

Fig. 2. Cumulative Distribution Figure of Traffic Volume

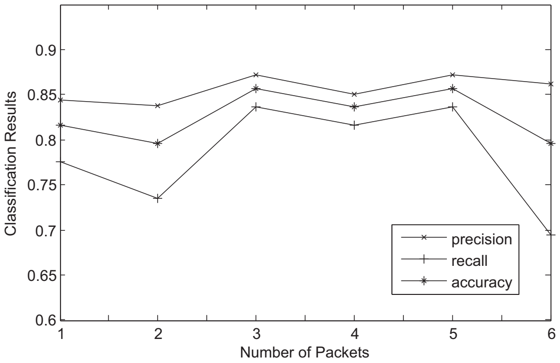

Fig. 3. Classification Results versus Different Number of Packets

IV. Implementation and Evaluation

A. Experimental Methods

1) Testbed Setup: The trace set we use is real-world traffic traces collected at a large institute (includes thousands of servers) in campus network. The trace sets consist of three 2 GB traces: trace-1, trace-2 and trace-3. All these traces were collected in different months in the year 2009. The evaluation platform is a generic PC, with a Intel(R) Core(TM)2 i3 CPU U380 (2 cores@1.33 GHz) and 2 GB DDR-II memory.

2) Evaluation Metrics: We define elephants as positive samples in this paper. To characterize the classifier's accuracy, we use common metrics known as *False Positive*, *False Negative*, *True Positive* and *True Negative*, whose definition are clearly described in [8]. Utilizing these concepts, we employed performance parameters: *Precision*, *Recall* and *Accuracy*, which are widely used in the area of Machine Learning. These metrics are defined as follows:

- *Precision*: Percentage of truly elephant samples among those classified as elephants.
- *Recall*: Percentage of elephant samples that are correctly classified as elephants.
- *Accuracy*: Percentage of the correctly classified samples among the total samples.

3) Classification Threshold: The threshold, which separates elephants from mice, is an important parameter and will dramatically influence the classification results. To generate an appropriate threshold, we give a cumulative distribution figure of traffic volume to help the determination.

Fig. 2 illustrates the cumulative distribution figure of traffic volume in different traces, and an obvious inflection point area is shown. Using numerical analysis tool MATLAB, we generate the fitting curve for each trace and calculate the mean value of those inflection points. Finally we treat this mean value as the threshold, which is 2.1 MB in our experiment.

Based on this threshold, we label each flow a mark of elephant or mouse. The number rate and byte rate of elephants in the traces are shown in TABLE III. Obviously, a very small percentage of flows, namely the elephants, are responsible for a very high percentage of the traffic volume, which shows a typical “elephants and mice phenomenon” in our test traces.

TABLE III
Elephants Statistics of Trace Set

Trace	Number of elephants	Bytes of elephants
trace-1	2.57%	89.76%
trace-2	2.27%	87.28%
trace-3	1.80%	85.90%

B. Evaluation Results

1) Feature Dimensions: As discussed above, number of packets, which represents different feature dimensions in the construction of SVM classifier, will influence the classification results. We use trace-1 as an example to calculate the variation of the classification results versus different packet numbers.

Fig. 3 shows that the metrics slightly increase when the number of packets grows from one to three. However, after a smooth tendency during the packet number changing from three to five, the classification results begin to decrease along the increasing of packet number. This looks implausible, but is overall the same in other experiment traces. The analysis of this phenomenon will be demonstrated later in section V.

2) Classification Results and Comparison: To get a preferable result, we choose the first three packets' size as features of SVM classifier, for the reason that a small packet number will result in fast training and testing speed, and the classification results remains accurate as well. Moreover, a large packet number will augment the latency. In summary, the most accurate and practical classification results can be achieved when three packets are used along with the 2.1 MB threshold.

For the reason that we do not find approaches similar to with Emilie, we compare Emilie with a simple dichotomy used by [14], whose original intention is close to that of Emilie. In this method, mice are composed less than or equal to 20 packets and elephants are those remain. Multiple times experiments shows a steady trend, and the average classification results shown in TABLE IV.

Precision Precisions of Emilie in different traces are over 87%, among which the highest one is over 90%. At the same time, precisions of the simple method are relatively poor, which are only around 3%. Using simple method will incur that large number of mice are misclassified as elephants and

TABLE IV
Precision, Recall and Accuracy of Trace Set

Trace	Method	Precision	Recall	Accuracy
trace-1	Emilie	87.27%	83.67%	85.71%
	Simple	3.84%	100.0%	37.87%
trace-2	Emilie	90.48%	80.85%	86.17%
	Simple	2.73%	100.0%	36.48%
trace-3	Emilie	88.00%	84.62%	86.54%
	Simple	3.55%	100.0%	38.17%

sent to fine-granular device to process, which will seriously impact performance.

Recall The inherent design of the simple method guarantees a zero misclassification rate for elephants, and Emilie also achieves efficient recalls which reach above 80% for all traces.

Accuracy The accuracy of Emilie fluctuates slightly around 86% with various traces, while that of the simple method is only about 37%.

Overall, we can achieve about 88% precision, about 85% recall and over 85% accuracy, which is sufficient for a practical traffic dispatcher. Meanwhile, though the simple method can achieve a perfect recall, it has poor precision and accuracy. So we conclude that Emilie is much more accuracy than the simple method.

Although we implement Emilie on a laptop PC for convenience, we achieve a classification speed at over 300 Kfps. And in our experiment, the flow size is about 400 KB on average, so we calculate that over 100 GBps throughput can be achieved, which is faster than most existing deep inspection devices. We believe that a remarkable improvement will be achieved if high performance server is used.

V. Discussion

In terms of precision, recall, accuracy and speed, the results above illustrate that Emilie achieves an outstanding performance as a traffic dispatcher. Besides, these results also imply several issues to be discussed.

The most interesting idea in this paper is that we only use sizes of the first several packets as features to determine whether a flow is an elephant or not. Inspired by the idea of [6] which focuses on the classification of application protocols, our originality is to achieve a predictable method to separate elephants from mice using sizes of the first several packets. The experimental results indicate that the Emilie is practical and efficient. The main reason is that various applications exchange different amount of interactive messages using diverse packet sizes between clients and servers at the beginning of communication. Meanwhile, the type of application has essential relationship with its flow size. Thus we can use packet sizes as features to classify. For example, the application of Video and FTP contribute to elephants for most cases, while the application of DNS and SMTP generally devote to mice.

Another interesting phenomenon is that the classification results deteriorate on average when over five packets are used. This can also be explained by the reason above. Applications exchange its private interactive messages through the first several packets, and after that, data transfer stage begins. So the employment of packets in data transfer stage may deprave the classification results.

VI. Conclusions and Future Work

In this paper we propose a novel traffic dispatching approach called Emilie, which utilizes Machine Learning techniques to classify elephants and mice. Based on the lengths of the first three packets of each flow, Emilie achieves an early identification of elephants and dramatically enhances the power of flow-based traffic identification techniques. Using the traffic trace datasets collected from real networks, we evaluate the performance of Emilie. The experimental results illustrate that on average about 88% precision, 85% recall and over 85% accuracy are achieved while predicting elephants.

In our future work, we will combine Emilie with counting based methods to integrate the prediction with precise revision, and then deploy this system in real network environment to further verify the practicality and efficiency.

VII. Acknowledgement

This work was supported by the National Key Technology R&D Program of China under Grant No.2012BAH46B04.

References

[1] C. Labovitz, S. Johnson, D. McPherson, J. Oberheide, and F. Jahanian, “Internet Inter-Domain Traffic,” in *Proceedings of the 2010 ACM SIGCOMM*, 2010.

[2] A. Callado, C. Kamienski, G. Szabo, B. Gero, J. Kelner, S. Fernandes, and D. Sadok, “A Survey on Internet Traffic Identification,” *IEEE Communications Surveys and Tutorials*, vol. 11, no. 3, pp. 37–52, 2009.

[3] T. Karagiannis, K. Papagiannaki, and M. Faloutsos, “BLINC: Multilevel Traffic Classification in the Dark,” in *Proceedings of the 2005 ACM SIGCOMM*, 2005.

[4] *L7-Filter*. [Online]. Available: http://l7-filter.clearfoundation.com/

[5] A. Moore and D. Zuev, “Internet Traffic Classification Using Bayesian Analysis Techniques,” in *Proceedings of the 2005 ACM SIGMETRICS*, 2005.

[6] B. Yang, G. Hou, L. Ruan, Y. Xue, and J. Li, “SMILER: Towards Practical Online Traffic Classification,” in *Proceedings of the 7th Architectures for Networking and Communications Systems*, 2011.

[7] *IANA Service Name and Transport Protocol Port Number Registry*. [Online]. Available: http://www.iana.org/assignments/service-names-port-numbers/service-names-port-numbers.xml

[8] T. Nguyen and G. Armitage, “A Survey of Techniques for Internet Traffic Classification using Machine Learning,” *IEEE Communications Surveys and Tutorials*, vol. 10, no. 4, pp. 56–76, 2008.

[9] W. Fang and L. Peterson, “Inter-AS Traffic Patterns and Their Implications,” in *Proceedings of the 1999 GLOBECOM*, 1999.

[10] A. Metwally, D. Agrawal, and A. Abbadi, “An Integrated Efficient Solution for Computing Frequent and Top-k Elements in Data Streams,” *ACM Transactions on Database Systems*, vol. 31, no. 3, pp. 1095–1133, 2006.

[11] M. Charikar, K. Chen, and M. Farach Colton, “Finding Frequent Items in Data Streams,” *Theoretical Computer Science*, vol. 312, no. 1, pp. 3–15, 2004.

[12] G. Manku and R. Motwani, “Approximate Frequency Counts over Data Streams,” in *Proceedings of the 28th International Conference on Very Large Data Bases*, 2002.

[13] J. Erman, A. Mahanti, and M. Arlitt, “Byte Me: A Case for Byte Accuracy in Traffic Classification,” in *Proceedings of the 3rd Annual ACM Workshop on Mining Network Data*, 2007.

[14] N. Azzouna and F. Guillemin, “Analysis of ADSL Traffic on an IP Backbone Link,” in *Proceedings of the 2003 GLOBECOM*, 2003.