Linking Network Usage Patterns to Traffic Gaussianity Fit

Ricardo de O. Schmidt*, Ramin Sadre‡, Nikolay Melnikov§, Jürgen Schönwälder§, Aiko Pras*
*University of Twente, The Netherlands
Email: {r.schmidt,a.pras}@utwente.nl
‡Aalborg University, Denmark
Email: rsadre@cs.aau.dk
§Jacobs University Bremen, Germany
Email: {n.melnikov,j.schonwalder}@jacobs-university.de

***Abstract*—Gaussian traffic models are widely used in the domain of network traffic modeling. The central assumption is that traffic aggregates are Gaussian distributed. Due to its importance, the Gaussian character of network traffic has been extensively assessed by researchers in the past years. In 2001, researchers showed that the property of Gaussianity can be disturbed by traffic bursts. However, assumptions on network infrastructure and traffic composition made by the authors back in 2001 are not consistent with those of today's networks.**

The goal of this paper is to study the impact of traffic bursts on the degree of Gaussianity of network traffic. We identify traffic bursts, uncover applications and hosts that generate them and, ultimately, relate these findings to the Gaussianity degree of the traffic expressed by a goodness-of-fit factor. In our analysis we use recent traffic captures from 2011 and 2012. Our results show that Gaussianity can be directly linked to the presence or absence of extreme traffic bursts. In addition, we also show that even in a more homogeneous network, where hosts have similar access speeds to the Internet, we can identify extreme traffic bursts that might compromise Gaussianity fit.

***Index Terms*—Traffic measurements, Gaussian modeling, Traffic analysis.**

I. Introduction

Traffic modeling is widely used for network planning, deployment and management. Models are used to identify and characterize traffic for purposes ranging from security to network dimensioning. Since the 90's, Gaussian traffic models have received special attention among researchers when studies revealed the presence of characteristics such as self-similarity and long-range dependence in modern network traffic [1], [2], [3]. It turned out that the fractional Brownian motion and other Gaussian models have many desirable properties for the modeling of IP traffic. The presence of long-range dependence and its long-term evolution were also studied in more recent studies, such as [4].

In this context, an important question is under what conditions network traffic can be assumed to be Gaussian. The Central Limit Theorem states that aggregated metrics, such as the amount of traffic transported per time unit on a network link, are normally distributed if a sufficiently large number of independent random variables are involved. Researchers have studied what a "sufficiently large number" could be. Previous works from 2002 [5] and 2006 [6] studied the amount of horizontal aggregation (*i.e.,* timescale for aggregating traffic) and vertical aggregation (*i.e.,* number of hosts and amount of transferred traffic) needed to justify that the traffic offered in an arbitrary timescale is Gaussian. In our own work [7], we showed, by performing tests for a very long measurement period of six years, that it is safer to relate high Gaussianity to traffic bandwidth than to the number of users.

A question quasi complementary to the above one was investigated in [8] in 2001. The authors studied why some traffic is *not* Gaussian. They showed that network traffic can be decomposed into a "beta" part which is nearly Gaussian and strongly long-range dependent and an "alpha" part which constitutes a small fraction of the total traffic and which is responsible for traffic bursts. The authors showed that generally very few high-rate connections dominate during a burst and reasoned that the majority of them are due to large file transfers over fast links.

With regard to modern network traffic, the work in [8] depicts several limitations. Being published in 2001, it assumes a strongly heterogeneous infrastructure where Ethernet lines and slow 56k modems coexist. In fact, the fasted alpha traffic flow was more than 50 times faster than the slowest beta traffic flow. In addition, although it provided a successful model of the network traffic, it did not study the *quantitative* relationship between the presence of bursts and the degree of Gaussianity.

Contribution. *The goal of this paper is to study the impact of traffic bursts on the degree of Gaussianity of network traffic using recent traffic measurements.* We apply the concept of alpha and beta traffic to an extensive set of traffic measurements from 2011 and 2012. We identify the traffic bursts and analyze the applications and hosts that generated them. Furthermore, we link the bursts to the degree of Gaussianity of the traffic, expressed as the goodness of fit. In our study we consider timescales of 100ms and 1s, which dominate users' perceived Quality of Service and, hence, are used for bandwidth provisioning approaches that often rely on Gaussian characteristic of traffic.

ISBN 978-3-901882-58-6

Organization. The remainder of this paper is organized as follows. In Sec. II we present the definition of Gaussianity used in this paper. In Sec. III we describe the network traffic datasets used in our experiments. In Sec. IV we describe the performed traffic analysis and the obtained results. We conclude the paper in Sec. V.

II. Gaussianity and Goodness of Fit

In this section we present the methodology we used to assess the Gaussianity of network traffic. To comply with recent related work, the study presented in this paper uses the same methodology from previous works [5], [6], [7].

A. Definition of Gaussianity

Consider $L_1(T), \ldots, L_n(T)$ to be the amount of traffic in bytes observed in time periods $1, 2, \ldots, n$ of length T, where $T > 0$ defines the timescale of traffic aggregation. The traffic aggregate $L(T)$ is Gaussian if it follows a normal distribution, *i.e.,* $L(T) \sim \text{Norm}(\rho, \sigma^2)$, where ρ is the average and σ^2 is the estimated variance of $L(T)$ given by, respectively

$$\rho = \frac{1}{n} \sum_{i=1}^{n} L_i(T)$$

and

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^{n} (L_i(T) - \rho)^2 \ .$$

B. Assessing Gaussianity of $L(T)$

Perhaps the most straightforward manner of assessing Gaussianity of samples is by means of quantile-quantile (Q-Q) plots. These plots provide a simple way to visually check whether a sample seems to be Gaussian or not. A Q-Q plot is created by plotting the inverse of the normal cumulative distribution function $\text{Norm}(\rho, \sigma^2)$ against the ordered statistics of the sampled data $L(T)$. Hence, the pairs for Q-Q plot are defined by

$$\left(\Phi^{-1} \left(\frac{i}{n+1} \right), \alpha_{(i)} \right), \ i = 1, 2, \ldots, n \ , \qquad (1)$$

where Φ^{-1} is the inverse of the normal cumulative distribution function, $\alpha_{(i)}$ are the ordered traffic averages of $L(T)$, for each time bin of length T, and n the size of our sample (*i.e.,* the number of time bins of size T). Note that in Eq. (1), for the inverse of the normal cumulative distribution function, the denominator $n+1$ is used instead of n because in normal distribution the 100th percentile is infinite. However, according to [9], [10], for large sample sizes (*i.e.,* large n) the difference of using one denominator or another is negligible.

As mentioned above, Q-Q plots provide a good visual way to check whether a sample seems to be Gaussian or not. To quantify the Gaussianity *goodness of fit* for a large amount of samples, though, a more scalable approach is needed. To comply with previous works [5], [6], [7] we use the *linear correlation coefficient*, defined in [11]. The *linear correlation coefficient* is determined by

$$\gamma(x, y) = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 \sum_{i=1}^{n} (y_i - \overline{y})^2}} \ , \qquad (2)$$

where the pair (x, y) is given by Eq. (1). That is, x is the inverse of the normal cumulative distribution function and y is the ordered statistics of traffic averages (*i.e.,* $\alpha_{(i)}$).

It was showed in [6] that if the *linear correlation coefficient* is relatively high (*e.g.,* $\gamma > 0.9$), then the hypothesis that the underlying distribution is normal distributed is equivalent to a *Kolmogorov-Smirnov* test for normality at significance 0.05. There are other tests for normality. For example, the authors of [12] extensively studied the performance of different tests and discussed their advantages and disadvantages for various situations (*e.g.,* when the empirical distribution is bimodal or long-tailed). In order to make the results of this paper comparable to those found in earlier publications, especially [6], we use γ to express the degree of Gaussanity. Whenever a trace has $\gamma \geq 0.9$ we refer to it as a "Gaussian trace" and, oppositely, a trace with $\gamma < 0.9$ is referred to as a "not Gaussian trace" (or non-Gaussian).

Given that we use a large amount of traffic traces in the experiments presented in this paper, we assess the Gaussianity *goodness of fit* of these traces by solely calculating the *linear correlation coefficient* for each trace individually. In the next sections we described the traffic datasets used in this analysis and present our experimental results.

III. Datasets and Traffic Characteristics

The datasets used in this work are comprised of packet-level traffic captures at three different links. These datasets comprise a total of 119 15-minute traces. Note that the trace duration of 15 minutes has been chosen in accordance with [6], [7], and that longer periods are generally not stationary due to the diurnal pattern. Moreover, our study focus on timescales of 100ms and 1s, which are of interest, for example, to bandwidth provisioning operations. Therefore, periods longer than 15 minutes might not be "stationary enough" for these timescales. These traces account for almost 30 hours of capture time and more than 7.7 billion packets. Table I summarizes the used datasets. Note that the column "length" gives the total capture duration of all, not necessarily consecutive, 15-minute traces for each location. Also note that, although Table I presents the average link use for each location, such value is generally not constant over the measurement period. For all locations throughput varies due to diurnal traffic patterns. We briefly describe the three measurement locations and present traffic characteristics in the following.

A. Datasets Locations

1) Dataset from location B: In location B the traces were measured in a 10 Gb/s up/down link at the core router of a university in the Netherlands. The link comprises all the incoming and outgoing traffic of the university. A total of approximately 886K IP addresses were observed during the

TABLE I
SUMMARY OF MEASUREMENTS

abbr.	description	year	length	# of hosts	link capacity	load
B	university border router in the Netherlands	2012	6h	886k	10 Gb/s	10%
C	university border router in Brazil	2012	18h45m	10.5k	155 and 40 Mb/s	19%
D	backbone links interconnecting two cities in USA	2011–12	5h	3M	2 × 10 Gb/s	10%

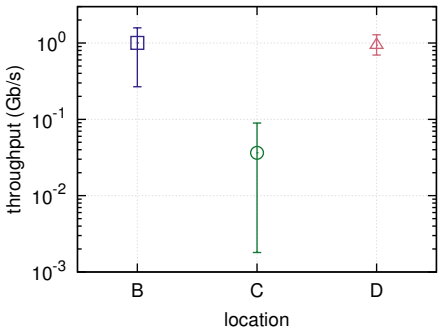
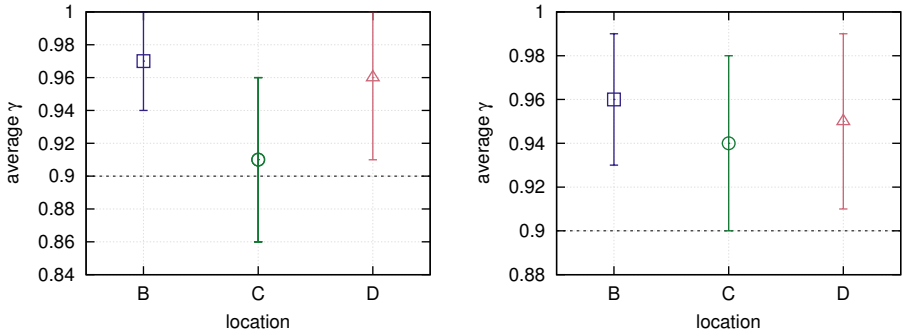
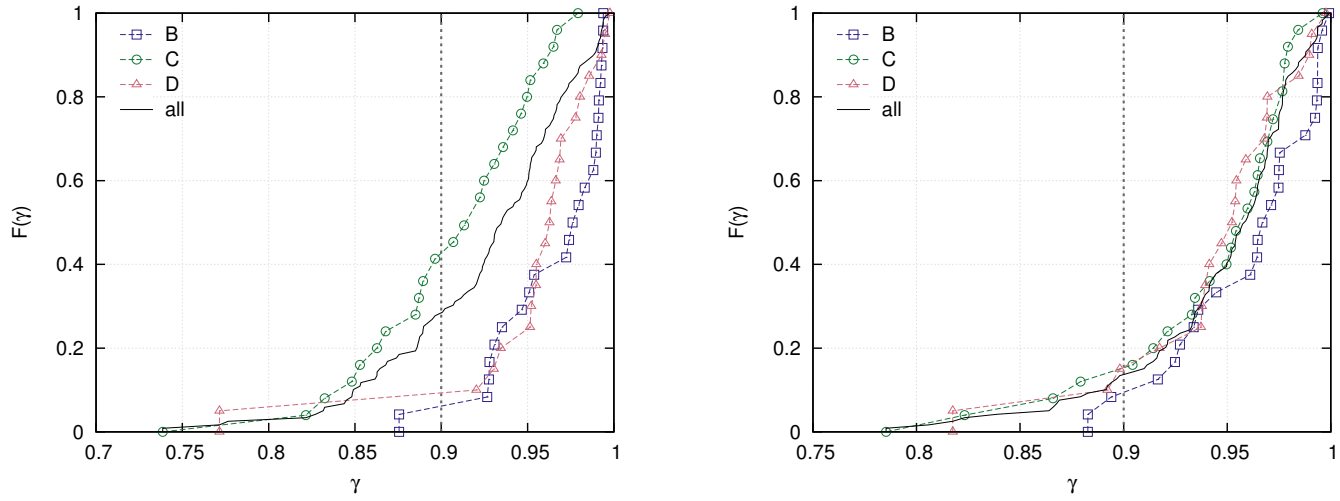


Fig. 1. Average traffic rate (throughput). Error bars show the minimum and maximum rate observed.



(a) $T = 100$ms (b) $T = 1$s

Fig. 2. Average γ for all traces in our datasets. Error bars show the standard deviation of γ.



(a) $T = 100$ms (b) $T = 1$s

Fig. 3. CDF of Gaussianity goodness of fit γ for all traces in our datasets.

measured period and they generated an average link use of 10% (up to 15% in busiest hours). This is a full day measurement in which traffic was captured during the first 15 minutes of every full hour for a period of 24 hours. Therefore, this location comprises a total of 24 15-minute traces.

2) Dataset from location C: The traces from location C were measured in the core router of a university in Brazil. The aggregate of two links of 155 Mb/s and 40 Mb/s was measured. It comprises traces collected during week days from September 2012 to December 2012. Each trace corresponds to the first 15 minutes of every full hour between 08:00 and 23:00 inclusive. Most of the traffic at location C is web browsing and e-mail.

3) Dataset from location D: Traces from dataset D are from the CAIDA public repository [13], [14] and were measured between December 2011 and February 2012. These traces were measured in a backbone link of a Tier1 ISP between the cities of San Jose and Los Angeles in the USA. The measured links have a total of 10 Gb/s each. Approximately 3 million unique IP addresses were observed during the monitored period. The links have an average load of 10%.

B. Link Usage

Table I presents the link load per location. However, this value is not constant overtime. Fig. 1 shows the average traffic

rate per 15-minute trace for each location. The figure also shows the minimum and maximum values of mean rate per trace. Traffic variations for location C are the lower ones, but this is also the location with the less active hosts and link capacity. Traffic of location B is the one that varies most. That can be explained by the fact that this is a 24-hour measurement and, therefore, low averages are most likely to be from the overnight period, while high averages are from the day. Variations of rate in location C can also be explained by the hour of the day the measurement took place. Although location C does not have measurements from the overnight period, low averages are likely to be from the day shifts (*e.g.,* 12:00–13:30 and 17:30–19:00) in which students and staff of the university are not actively using the network. Traffic from location D is more stable traffic, as shown in Fig. 1. That might be expected since for this location measurements are months apart, but happened always during the same hour of the day.

C. Traffic Gaussianity

Using the methodology described in the previous section, we have computed γ for all traces in our datasets. Fig. 2 shows the average and standard deviation of γ for all traces per location. As one can see, traffic in our datasets tends to be Gaussian, *i.e.,* all averages are above the threshold $\gamma > 0.9$. Furthermore, one can see that the difference on average γ does not crucially change from $T = 100$ms (Fig. 2a) to $T = 1$s (Fig. 2b). However, averages may be misleading because for all locations we do have few traces with $\gamma < 0.9$ and, therefore, are not Gaussian.

In order to understand to which extend our traces are Gaussian, and also to support the Gaussianity analysis in the following section, Fig. 3 shows the cumulative density function (CDF) of Gaussianity goodness of fit γ for all traces in our datasets. For $T = 100$ms, roughly 28% of all traces are not Gaussian. Clearly, most of them belong to location C, where around 40% of traces have $\gamma \leq 0.9$. For $T = 1$s the number of non-Gaussian traces is reduced to around 13% of all traces in our datasets. The majority of those is still from location C, where roughly 15% of traces are not Gaussian. The difference on Gaussianity fit for traces from location C between $T = 100$ms and $T = 1$s can be explained by the lower traffic rate in the measured links for this location. By reducing the size of the time bins (*i.e.,* T), the small amount of traffic from traces of location C is aggregated in a few bins, interspersed with empty bins without any network packets, ultimately resulting in an on/off-like behavior. As already shown in previous works [5], [7], this behavior considerably reduces the Gaussianity fit.

However, as one observes in Fig. 3, the opposite situation can also happen: some traces have poorer Gaussianity fit at larger T. This behavior has also been briefly explained in [7]. For such traces, traffic bursts at smaller timescales (*e.g.,* $T = 100$ms) are very close to each other in time. Therefore, once we aggregate this traffic into larger bins (*e.g.,* $T = 1$s), traffic of these bursts are grouped into few bins, resulting in bursts that reach much higher rates than the average rate of the trace. In the end, these few bins with exceptionally high rates disturb the Gaussianity fit at larger T.

For a thorough study of the Gaussianity fit of the datasets used in this work, we refer to [7]. In the next section our analysis focuses on the differences between traces with low and high Gaussianity fit.

IV. Traffic Analysis

In this section we present a thorough analysis of traffic characteristics and their potential relationship to Gaussianity. We start by demonstrating the impact of bursts on the Gaussianity fit. Then, we show that these bursts are mostly related to single applications running in the network, and we also assess the impact of individual hosts on Gaussianity. Finally, we argue why identifying applications and linking them to the degree of Gaussianity of traffic is a quite complex and challenging task.

A. Impact of Bursts on Gaussianity

Normal distributed traffic is expected to have bursty and calm moments. By definition, if the traffic aggregate $L_i(T)$ follows a normal distribution Norm(ρ, σ^2), the probability that it exceeds x is given by the complementary CDF

$$P(L_i(T) > x) = 1 - \frac{1}{2}\left(1 + \text{erf}\left(\frac{x-\rho}{\sqrt{2\sigma^2}}\right)\right) . \quad (3)$$

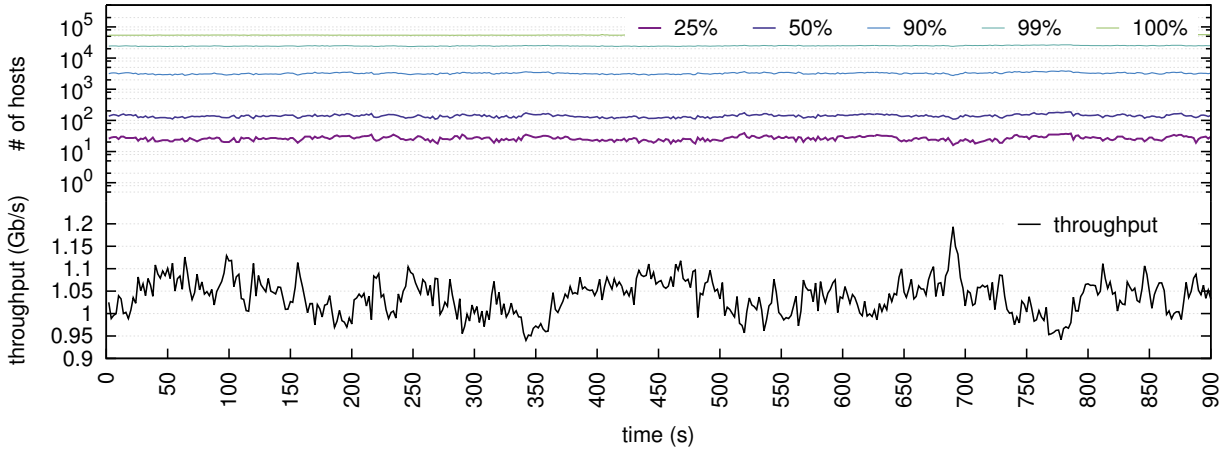
Fig. 4 shows the difference between two traces from location D with low and high Gaussianity fit, respectively. The curve in the bottom half of these plots shows the traffic aggregate of the sample traces over the measurement period of 15 minutes. Note that we have chosen $T = 2$s in this figure for visualization purposes, while we use $T = 100$ms and $T = 1$s in all other experiments. The trace from Fig. 4a has one of the highest Gaussianity fits in our datasets ($\gamma = 0.9977$). One can clearly see that traffic of this trace has regular ups and downs and in any moment a burst really protrudes from the baseline traffic. Contrariwise, Fig. 4b shows the traffic time series of one of the traces with the lowest degree of Gaussianity ($\gamma = 0.8175$) among all traces we have collected. In this case, one can easily notice the very high bursts during the time period 50–200 and at time 420.

We have manually inspected and compared several traces with poor and good Gaussianity and noticed that such bursts are typical for poorly Gaussian traces. In order to assess this behavior systematically, we define a burst as a time bin where the traffic aggregate exceeds the threshold θ defined by

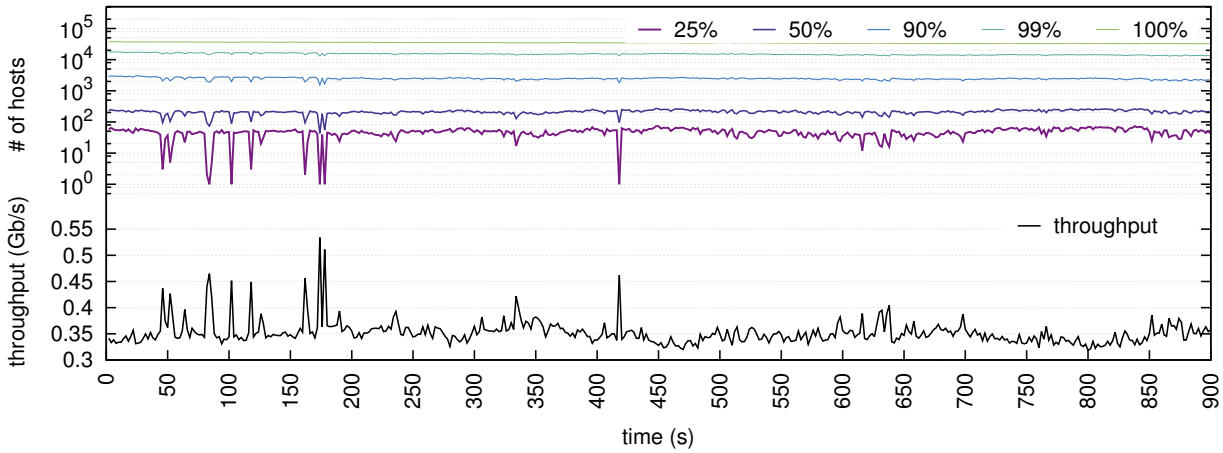
$$\theta = \rho + 3\sigma . \quad (4)$$

That is, θ is three standard deviations above the trace average rate ρ. A similar definition of burstiness has been used in [8], [15]. According to Eq. (3) this should only happen with probability 0.00135 in perfect Gaussian traffic.

The plots of Fig. 5 and 6 show for each trace of different locations, respectively at $T = 100$ms and $T = 1$s, its Gaussian fit γ and the percentage of time bins that exceed the above



(a) Good Gaussian trace ($\gamma = 0.9977$)



(b) Bad Gaussian trace ($\gamma = 0.8175$)

Fig. 4. Traffic aggregate and traffic shares at $T = 2$s for two sample traces from location D.

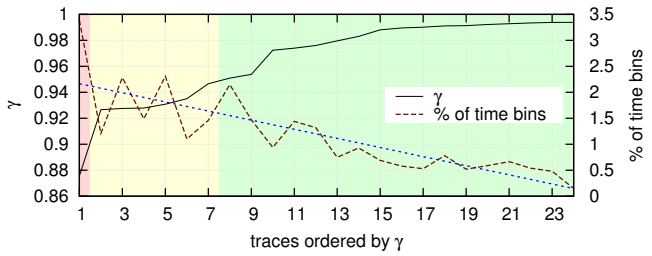
threshold θ. In these plots, traces are sorted by their Gaussian fit, *i.e.,* trace 1 (left on the x-axis) is the trace with the lowest γ. Note that the traces positioning in the x-axis varies from $T = 100$ms to $T = 1$s due to their different values of γ at different T. That is, trace 1 in Fig. 5a may not be the same as trace 1 in Fig. 6a. Moreover, the colors of the background in these plots indicate the regions where $\gamma < 0.9$, $\gamma < 0.95$ and $\gamma \geq 0.95$, respectively. These considerations are also valid for plots from Fig. 8 to 11.

Although the resulting curves in these pltos depict strong fluctuations independently of T, we observe an inverse relationship between the amount of bursts exceeding the threshold and the Gaussian fit. That is, non-Gaussian traces tend to have more bursts than Gaussian ones. This tendency is highlighted by the least-squares-fitted diagonal dotted line. Note that this is not a trivial outcome since non-Gaussianity could be also caused by the *absence* of bursts. In fact, a few non-Gaussian traces have a very small number of bursts.

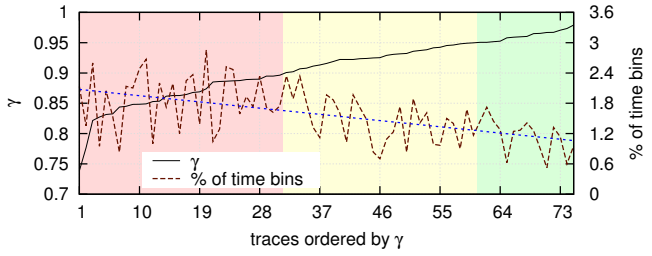
B. Impact of Applications on Gaussianity

In the previous section we have shown the relationship between bursts and (non-)Gaussianity. In this section, we want to study which applications are responsible for bursts. Note that we use the straightforward port-matching method for identifying applications. However, we are aware of the drawbacks of such method. Challenges related to traffic classification and its connection to Gaussianity are further discussed in Section IV-D.

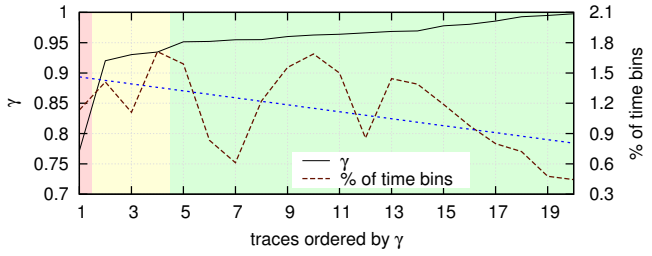
Fig. 7 shows the traffic aggregates of three sample traces (one from each location) with a low Gaussianity fit. The upper curve gives the aggregate of all traffic. We observe several bursts. For the trace from location B, the lower curve only shows the aggregate for traffic transferred on port 563 (*i.e.,* NNTP). For the sample traces from locations C and D, the lower curve shows the aggregate of traffic transferred on ports 80 and 443 (*i.e.,* HTTP and HTTPS, respectively). It can be seen that the protocol-specific curves follow closely the shape of the bursts for all the three examples. In fact, we have observed that typically a burst consists entirely of traffic



(a) Traces from B

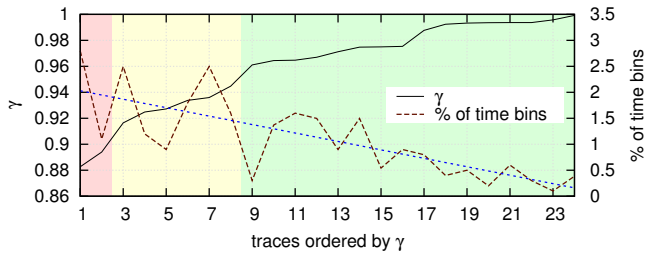


(b) Traces from C

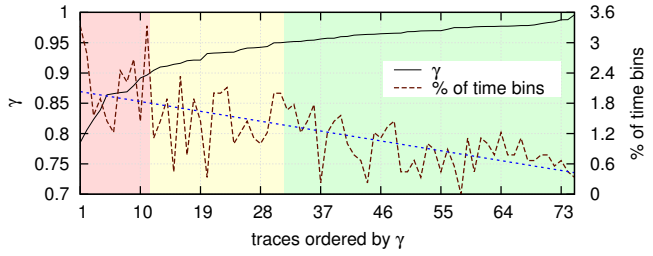


(c) Traces from D

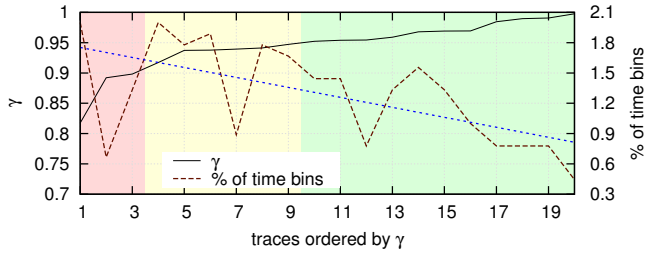
Fig. 5. Percentage of time bins with bursts at $T = 100$ms.



(a) Traces from B

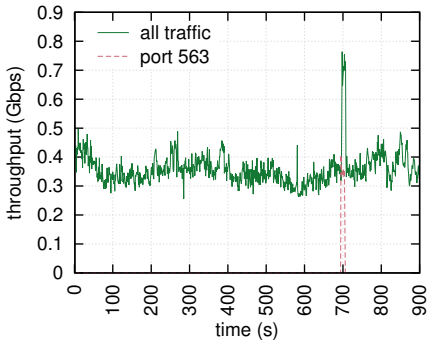


(b) Traces from C

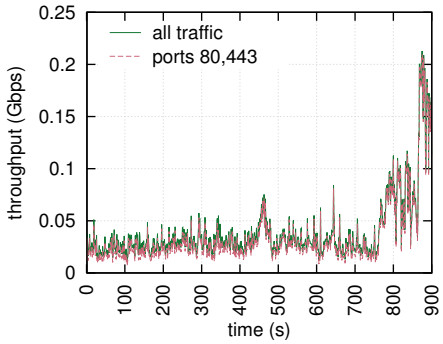


(c) Traces from D

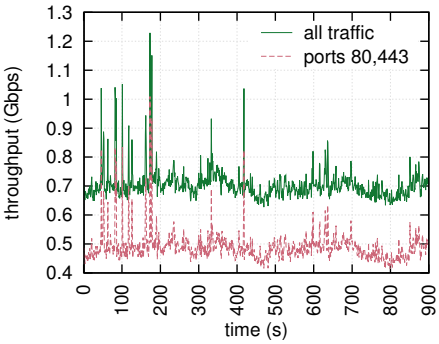
Fig. 6. Percentage of time bins with bursts at $T = 1$s.



(a) Trace from B



(b) Trace from C

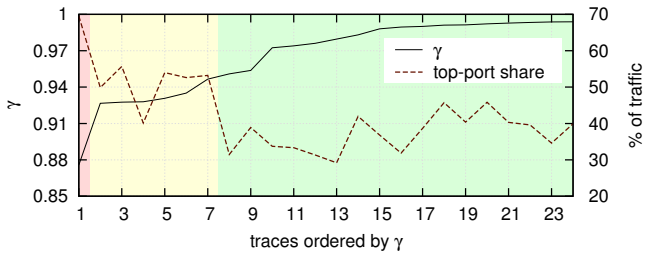


(c) Trace from D

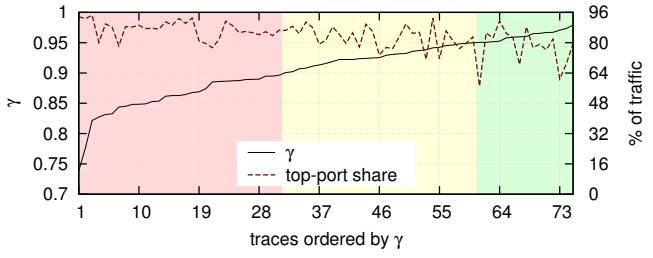
Fig. 7. Traffic aggregates at $T = 1$s and the port causing the bursts.

from only one application and, hence, removing the specific traffic of such application from the trace would also remove the bursts. Note that all bursts in a trace are not necessarily caused by the same application.

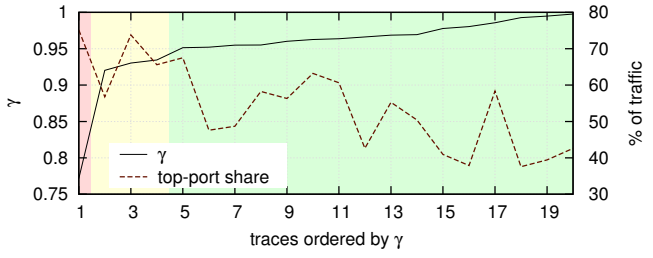
In order to validate this observation for the entire dataset, we have calculated for each burst that exceeds θ (as defined in Eq. 4) the share of the traffic on the most active port in the time bin of that burst, and computed an average share for all bursts of a trace. The plots in Fig. 8 and 9 show the resulting (average) traffic share, for $T = 100$ms and $T = 1s$ respectively. Again, traces are sorted on the x-axis by their respective Gaussianity fit γ and the background color indicates the regions where $\gamma < 0.9$, $\gamma < 0.95$ and $\gamma \geq 0.95$, respectively. We observe that for traces with low γ, the traffic bursts that exceed θ mainly consist of traffic from the most active ports, and that this relationship weakens with increasing γ. Note that the share never reaches 100%. Clearly, this is because the time bin containing the burst also contains normal



(a) Traces from B

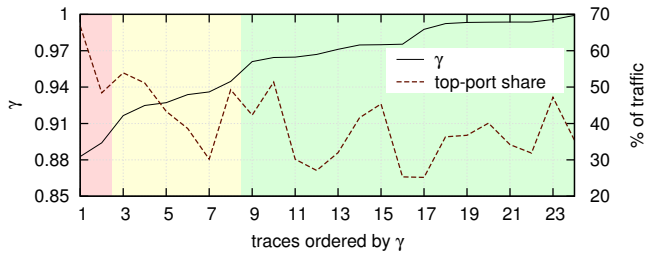


(b) Traces from C

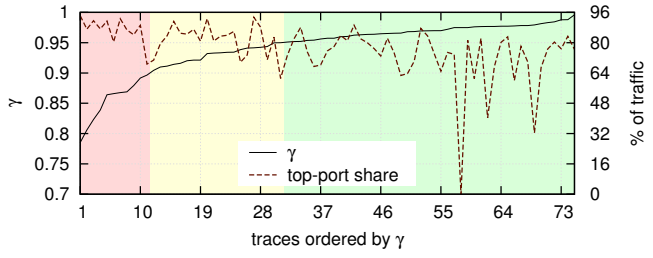


(c) Traces from D

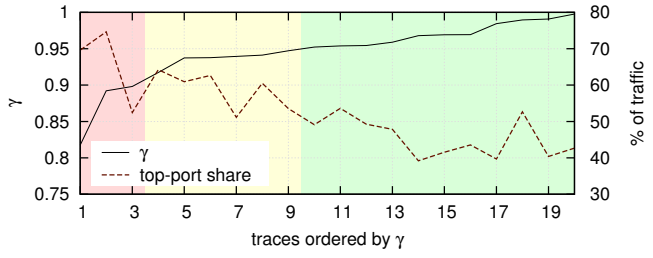
Fig. 8. Share of the most active applications in bursts at $T = 100$ms. Note the different scales of the y-axes.



(a) Traces from B



(b) Traces from C



(c) Traces from D

Fig. 9. Share of the most active applications in bursts at $T = 1$s. Note the different scales of the y-axes.

baseline traffic. Furthermore, we observe a generally high share for all traces of location C. This is because HTTP(S) is the most dominant traffic at this location.

C. Impact of Individual Hosts On Gaussianity

The previous analysis has shown that bursts are caused by single applications. In this section we investigate how individual hosts contribute to the traffic in such bursts.

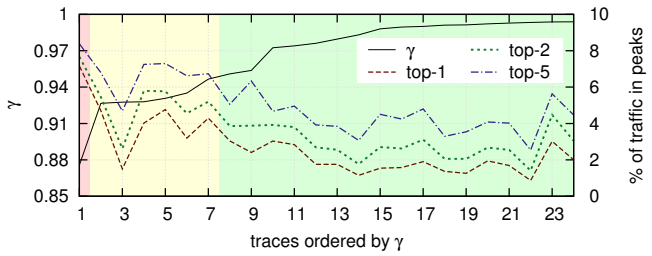
The five curves in the top half of the plots in Fig. 4 show the absolute number of the most active hosts that are responsible for 25%, 50%, 90%, 99%, and 100% of the traffic sent in a given time bin. More formally, let $b_1(t) \geq b_2(t) \geq \ldots$ be the sorted number of bytes sent by the hosts in time bin t, *i.e.,* $b_1(t)$ is the number of bytes sent by the most active host in time bin t, $b_2(t)$ is the number of bytes sent by the second most active host etc. The number $q_s(t)$ of the most active hosts responsible for a share s of the traffic in time bin t is defined as

$$q_s(t) = \min_{\sum_{i=1}^{x} b_i \geq s \cdot B(t)} x, \tag{5}$$

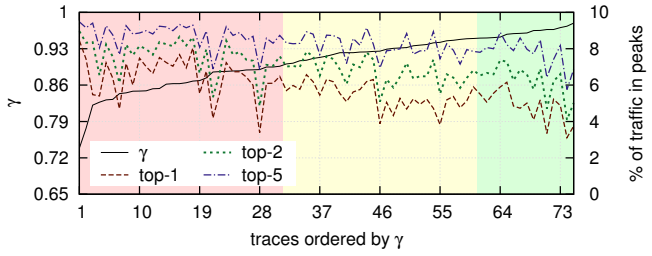
where $B(t)$ is the total number of bytes sent in time bin t.

One can see that while for the good Gaussian trace in Fig. 4a the number of users that are responsible for any share of the traffic remains quite constant over time. In any moment the number of contributing hosts drops considerably, not even during the highest traffic burst of this example trace, around 690s. On the contrary, for the trace with bad Gaussianity fit in Fig. 4b, the number of hosts that contribute to a certain share of traffic significantly drops during bursts. For example, during the burst at time 420s, only one host sends 25% of the traffic, which more or less corresponds to the difference between the 0.45 Gb/s peak throughput of the burst and the baseline throughput of 0.35 Gb/s, *i.e.,* that burst is caused by traffic from one single host. Outside the bursts, a much larger number of hosts contribute to the 25% traffic share. In general, the non-bursty part of the traffic is in accordance with the observations made in [16] that typically 90–95% of IP traffic is generated by 10–5% of the sources. The authors of [16] also found bursts in their traffic but mostly connected them to attacks. However, after a manual inspection of several bursts, we consider it unlikely that the bursts in our traces are caused by malicious activities. The size of the bursts also makes them very unlikely to be caused by source-level bursts on packet level [17].

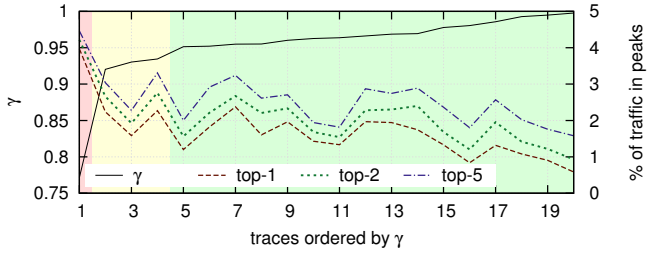
Again, we have validated this observation for the entire dataset. We have calculated for each burst higher than θ the



(a) Traces from B

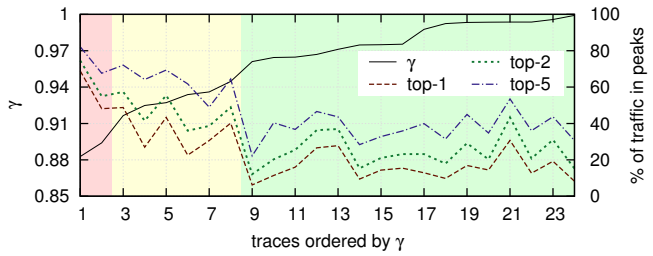


(b) Traces from C

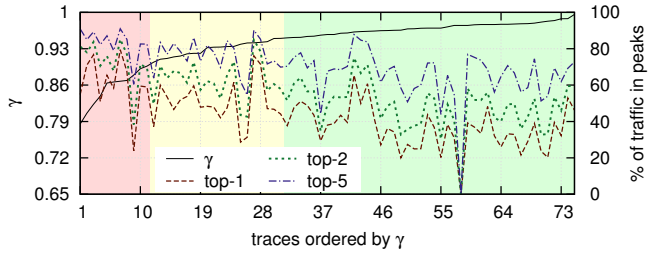


(c) Traces from D

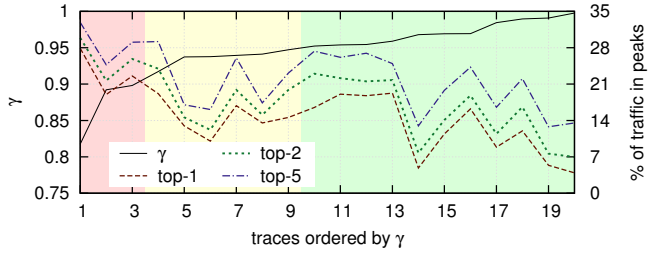
Fig. 10. Share of traffic for top-IPs in bursts at $T = 100$ms.



(a) Traces from B



(b) Traces from C



(c) Traces from D

Fig. 11. Share of traffic for top-IPs in bursts at $T = 1$s.

share of the most active host, the two most active hosts, and the five most active hosts to the traffic in that burst and computed the average shares for all bursts of a trace. The plots in Fig. 10 and 11 show the resulting traffic shares for each trace per location at $T = 100$ms and $T = 1$s, respectively. We observe for all traces that, independently of the Gaussianity fit, very few hosts are responsible for a significant amount of transferred traffic during the bursts. In some situations, at $T = 1$s, less than 5 hosts are responsible for more than 80% of all burst traffic in traces from location B and C and more than 35% in traces of D.

D. Challenges on Traffic Classification

In this paper we have identified applications to the level of protocol, *e.g.,* HTTP(S), by matching the port numbers, *e.g.,* 80 and 443. However, it is known that a plethora of applications are currently running on top of HTTP(S) and identifying those is not a straightforward task. Researches such as [18], [19] point out that the high rate of development of new applications and "bad habits" on implementing their communication blocks make traffic classification quite challenging. For example, many applications do not have IANA registered ports. Instead, they make use of well-known ports or tunneling to prevent detection and deceive filtering or firewalls. For instance, BitTorrent can also make use of random ports, complicating their identification by default communication ports.

Sophisticated traffic classification almost precludes application identification with goals of Gaussianity assessment. That is, to ultimately have a kind of *rule-of-thumb* that would allow us to make assumptions on the degree of Gaussianity of a given traffic aggregate based on the mix of applications found within it seems to be more complex than simply measuring the traffic for a short period and performing the same operations as we have done in this paper (*i.e.,* computing Gaussianity goodness-of-fit γ). Nonetheless, we have shown that bursts of traffic tend to belong to a handful of hosts and being transferred on a very limited range of ports, even for large networks. Hence, we also see our work as a first contribution toward an application-oriented method to assess the Gaussianity assumption: instead of performing a costly network-wide measurement on packet-level, researchers or network operators would first identify hosts that contribute most to bursts in the main traffic and later explore further the applications that are being used by those hosts (*i.e.,* to a higher level than the application protocol that they use).

V. Conclusion

In this paper, we have shown by an extensive analysis of recent network measurements that the degree of Gaussianity of network traffic, expressed by a goodness-of-fit factor, is directly linked to the presence of extreme traffic bursts. While fairly Gaussian network traffic is mostly burst free, traffic with a low Gaussian fit is during up to 3.6% of its duration bursty. Furthermore, we have shown that these bursts are mostly created by single applications. In particular, traffic bursts at two of our measurement locations mostly consist of HTTP(S) traffic. We have also observed that bursts in traffic with a low Gaussian fit tend to consist of traffic from only one application. Finally, we have shown that the traffic inside bursts is sent from only a few hosts. Our results allow the conclusion that poor Gaussianity is caused by short but intensive activities of single network hosts. This suggests that the bursts are related to transfers of big files over fast links.

Our findings confirm that the concept of alpha and beta traffic introduced by [8] in 2001 is still valid in recent network traffic. However, it is worth to note that two of our measurement locations were university core routers that connect a rather homogeneous set of hosts with identical or at least similar link speeds to the Internet. While the authors of [8] speculated that the diversity of clients could be a reason for the existence of alpha and beta traffic, our results indicate that the cause can probably be found in the characteristics of the servers. We plan to study this aspect in future work.

It is important to keep in mind that, although measurements used in this work have very distinct users and traffic nature, our dataset is definitely not fully representative of the whole Internet traffic. Nonetheless, our conclusions certainly pave the way for additional research in this area.

Acknowledgements

This work has been partially funded by Flamingo, a Network of Excellence project (#318488) and by UniverSelf project (#257513), both supported by the European Commission under its Seventh Framework Programme.

References

[1] 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*, vol. 2, no. 1, pp. 1–15, 1994.

[2] V. Paxson and S. Floyd, "Wide-Area Traffic: The Failure of Poisson Modeling," *IEEE/ACM Transactions on Networking*, vol. 3, no. 3, pp. 226–244, 1995.

[3] I. Norros, "A storage model with self-similar input," *Queueing Systems*, vol. 16, no. 3–4, pp. 387–396, 1994.

[4] P. Borgnat, G. Dewaele, K. Fukuda, P. Abry, and K. Cho, "Seven years and one day: Sketching the evolution of internet traffic," in *Proceedings of IEEE INFOCOM 2009*, 2009, pp. 711–719.

[5] J. Kilpi and I. Norros, "Testing the Gaussian approximation of aggregate traffic," in *Proceedings of the 2nd ACM SIGCOMM Internet Measurement Workshop*, ser. IMW'02, 2002, pp. 49–61.

[6] R. van de Meent, M. Mandjes, and A. Pras, "Gaussian Traffic Everywhere?" in *Proceedings of the of the IEEE International Conference in Communications*, ser. ICC'06, 2006, pp. 573–578.

[7] R. de O. Schmidt, R. Sadre, and A. Pras, "Gaussian Traffic Revisited," in *Proceedings of the 12th IFIP Networking Conference*, ser. Networking'13, 2013, pp. 1–9.

[8] S. Sarvotham, R. Riedi, and R. Baraniuk, "Connection-level Analysis and Modeling of Network Traffic," in *Proceedings of the 1st ACM SIGCOMM Workshop on Internet Measurement*, ser. IMW'01, 2001, pp. 99–103.

[9] L. Makkonen, "Bringing Closure to the Plotting Position Controversy," vol. 37, no. 3, 2008, pp. 460–467.

[10] L. Makkonen, M. Pajari, and M. Tikanmäki, "Closure to "Problems in the extreme values analysis"," vol. 40, no. 1, 2013, pp. 65–67.

[11] B. M. Brown and T. P. Hettmansperger, "Normal Scores, Normal Plots and Tests for Normality," *Journal of the American Statistical Association*, vol. 91, no. 436, pp. 1668–1675, 1996.

[12] T. Thadewald and H. Büning, "Jarque-Bera test and its competitors for testing normality: A power comparison," ECONSTOR, http://hdl.handle.net/10419/49919, Free University Berlin, School of Business & Economics, Tech. Rep. 2004/9, 2004.

[13] CAIDA, "The CAIDA UCSD Anonymized Internet Traces 2011 – Dec. 22," http://www.caida.org/data/passive/passive_2011_dataset.xml, online. Accessed Nov. 2013.

[14] ——, "The CAIDA UCSD Anonymized Internet Traces 2012 – Jan. 19 and Feb. 16," http://www.caida.org/data/passive/passive_2012_dataset.xml, online. Accessed Nov. 2013.

[15] K. Lan and J. Heidemann, "A Measurement Study of Correlation of Internet Flow Characteristics," *Computer Networks*, vol. 50, no. 1, pp. 46–62, 2006.

[16] A. Broido, Y. Hyun, R. Gao, and kc claffy, "Their Share: Diversity and Disparity in IP Traffic," in *Proceedings of the 5th International Passive and Active Network Measurement Workshop*, ser. PAM'04, 2004, pp. 113–125.

[17] H. Jiang and C. Drovolis, "Source-Level IP Packet Bursts: Causes and Effects," in *Proceedings of the the 3rd ACM SIGCOMM Conference on Internet Measurement*, ser. IMC'03, 2003, pp. 301–306.

[18] A. Tongaonkar, R. Keralapura, and A. Nucci, "Challenges in Network Application Identification," in *Proceedings of the 5th USENIX Conference on Large-Scale Exploits and Emergent Threats*, ser. LEET'12, 2012, pp. 1–1.

[19] A. Dainotti, A. Pescapé, and K. C. Claffy, "Issues and Future Directions in Traffic Classification," *IEEE Network*, vol. 26, no. 1, pp. 35–40, 2012.