Gaussian Traffic Revisited

Ricardo de O. Schmidt*, Ramin Sadre†, Aiko Pras*
*University of Twente, The Netherlands
Email: {r.schmidt, a.pras}@utwente.nl
†Aalborg University, Denmark
Email: rsadre@cs.aau.dk

***Abstract*—The assumption of Gaussian traffic is widely used in network modeling and planning. Due to its importance, researchers have repeatedly studied the Gaussian character of traffic aggregates. However, dedicated studies on this subject date back to 2002 and 2006. It is well known that network traffic has changed in the past few years due the the increasing use of social networks, clouds and video streaming websites.**

Therefore, the goal of this paper is to verify whether the Gaussianity assumption still holds for current network traffic. To this end, we study the characteristics of a large dataset, consisting of traces from four continents. The employed analysis methodology is similar to that found in previous works. In addition to the analysis of recent measurements, we also perform tests for a very long measurement period of six years. Our results show that the evolution of network traffic has not had a significant impact on its Gaussian character. Our findings also indicate that it is safer to relate the degree of Gaussianity to traffic bandwidth than to the number of users for high-speed links.

I. Introduction

Traffic modeling is very important for network design, planning, deployment and management. Models are used to identify and characterize traffic for purposes ranging from security to network dimensioning. Since the 1990s, Gaussian traffic models have received special attention when studies revealed the presence of characteristics such as self-similarity and long-range dependency 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 Gaussianity of aggregated network traffic is suggested by the Central Limit Theorem. In 2002, the authors of [4] studied the level of aggregation that is needed to justify the assumption that the amount of traffic offered in an arbitrary timescale follows a normal distribution. The follow-up work in [5] (2006) further explored the potential and limitations of Gaussian models with respect to the aggregation level.

Compared to six years ago, a much wider range of network applications can be found nowadays. Social networks, cloud computing and video streaming have drastically changed user behavior. There is, indeed, a clear evolution of traffic, as discussed in [15]. This raises the question whether, and to what degree, the assumption of Gaussianity still holds in today's networks. Intuitively, one would expect that the degree of Gaussianity has even increased since current networks generally aggregate traffic of a larger number of users/hosts than in the past. However, the measurements from [5] date back to 2004 or earlier and no recent results are available.

The goal of this paper is to verify the assumption of traffic Gaussianity by recent measurements. The methodology of the study presented in this paper borrows from [4], [5]. We assess the impact of different horizontal aggregation level, i.e., the chosen timescale of aggregation on Gaussianity, as well as the impact of the vertical aggregation level in terms of the number of hosts and the amount of traffic aggregated.

We show, by analyzing an extensive dataset of recent measurements, that the assumption of Gaussianity still holds for current network traffic. In contrast to existing work, we also perform tests for a very long measurement period of six years. Our results indicate that the evolution of network traffic has not had a significant impact on its Gaussian character. Our findings also suggest that it is safer to relate the degree of Gaussianity to traffic bandwidth than to the number of users for high-speed links.

The reminder of this paper is organized as follows. For our experiments, we have used traffic measured at several locations around the globe. The traffic traces are presented in Section II. In Section III we describe the methodology we follow to assess the Gaussian character of traffic. The results are presented in Section IV. Finally, we draw our conclusions in Section V.

II. Measurements Data Set

In this section we describe the measurement dataset used to assess the Gaussianity of network traffic. The entire dataset comprises 768 15-minute traces[1], totalling 192 hours of captures. These traces come from different locations around the globe and account for a total of more than 18.5 billion packets. Traffic captures were done at the IP packet level, using tools such as `tcpdump`. Table I gives a summary of the data obtained from the six measurement locations. Note that the column "length" gives the total

[1]The trace duration of 15 minutes has been chosen in accordance with [5]. Longer time periods are generally not stationary due to the diurnal pattern.

TABLE I
SUMMARY OF MEASUREMENTS

abbr.	description	year	length	# of hosts	link capacity	avg. use
A	link from university's building to core router	2011	24h	6.5k	2 × 1 Gb/s	15%
B	core router of university in the Netherlands	2012	6h	886k	10 Gb/s	10%
C1	core router of university in Brazil	2012	85h	10.5k	155 and 40 Mb/s	19%
C2	core router of university in Brazil (24-hour)	2012	6h	7k	155 and 40 Mb/s	11%
D	backbone links connecting Chicago and Seattle	2011	4h	1.8M	2 × 10 Gb/s	8%
E	backbone links connecting San Jose and Los Angeles	2011–2012	10h	3M	2 × 10 Gb/s	10%
F1	trans-Pacific backbone link	2012	13h	4M	n/a	n/a
F2	trans-Pacific backbone link (historical traces)	2006–2012	44h	1M	n/a	n/a

duration of the, not necessarily consecutive, 15-minute traces, i.e., a length of 1h corresponds to four traces.

A. Measurement locations

Location A: Location A is an aggregated link (2 × 1 Gb/s) connecting a university building in the Netherlands to the university's core router (university's gateway). Considering incoming and outgoing traffic, this link aggregates traffic from approximately 6500 hosts and has an average use of 15%. Most traffic in this link is actually internal to the university, i.e., from that building to other parts of the campus. Due to the small number of hosts, single activities, such as an overnight automatic backup, can drastically change the shape of the traffic. The measurement took place in a week day in September of 2011 with a duration of 24 hours. Therefore, this location comprises 96 successive 15-minute traces.

Location B: Location B is the 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 886000 IP addresses were observed during the 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.

Location C: Location C is the core router of a university in Brazil. The aggregate of two links of 155 Mb/s and 40 Mb/s was measured over two periods of time. $C1$ comprises traces collected during some week days from September 2012 to December 2012. Each trace corresponds to the first 15 minutes of each full hour between 08:00 and 23:00 inclusive. $C2$ traces are from the same monitored point as $C1$. However, $C2$ is a 24-hour measurement that contains the first 15 minutes of traffic of every full hour from 00:00 to 23:00 inclusive. The $C2$ measurement took place in December 2012. Most of the traffic at location C is web browsing and email.

Locations D and E: The traces for location D and E are from CAIDA's public repository [8], [9]. Two unidirectional backbone links of 10 Gb/s were measured for each location. The original traces are captures of a full hour done on selected days. In location D, links interconnecting Chicago and Seattle (USA) were measured and the selected traces are from May and July 2011. In location E, links interconnecting Los Angeles and San Jose (USA) were measured and the selected traces are from December 2011 and January and February 2012. Each full hour of capture gives us 4 successive 15-minute traces.

Location F: Location F is a trans-pacific link. Measurements for this location come from the public MAWI repository [10]. We do not have additional information on the link capacity and, consequently, we cannot determine the average use of the link[2]. Traces from this location are divided in two groups. $F1$ consists of very recent traffic captures, dating from November 2012 to December 2012. $F2$ comprises historical captures starting in August 2006 until December 2012. $F1$ traces aggregate traffic from an average of more than 4 million hosts, and $F2$ traces from an average of more than 1 million hosts. Note that for $F2$ traces, as we further show, traces dating from 2011 or older have much less hosts than 2012 traces. Unfortunately, MAWI does not provide additional details about the measurement point that could lead us to a better conclusion on the causes for such increase in the number of hosts in the year of 2012.

For measurements directly performed by us (i.e., locations A, B and C), no packet losses were observed. From CAIDA's web page, we know that for one link of the location D's pair, packet losses are likely to happen. For traces from location F, no information on packet loss is provided in the online repository.

B. Link usage

Although Table I presents the average link use for each location, such value is generally not constant over the measurement period. In fact, for some locations it varies quite a lot. Fig. 1 shows the average traffic rate per 15-minute for each location. The figure also shows the minimum and maximum values of mean rate per trace. As one can see, traffic from locations with higher-capacity links are the ones that vary most. In case of locations A and B, the difference between minimum and maximum values

[2]The information on the link capacity given on the MAWI website is not consistent with the bandwidth observed in the traces.

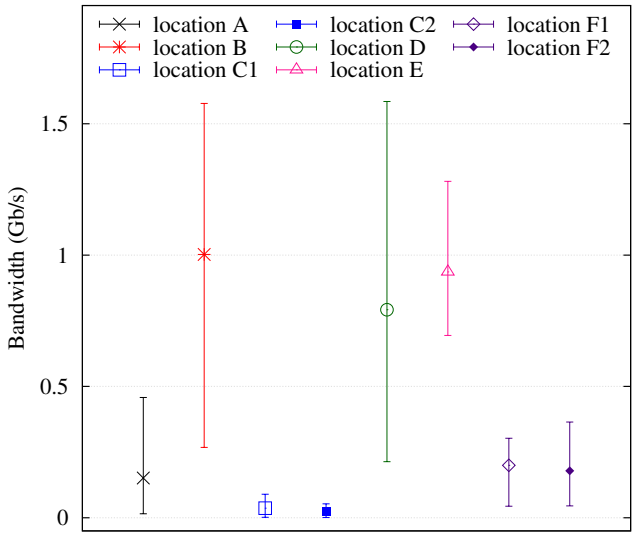


Fig. 1. Average, minimum and maximum values of mean traffic at each location; location C corresponds to $C1$ and $C2$ together, and location F corresponds to $F1$ and $F2$ together.

are due to the fact that these are 24-hour measurements. Therefore, low averages are most likely to be from the overnight period, while high averages from the day.

III. Methodology

In this section we present our procedure to study the Gaussianity characteristic of the traffic. The methodology of the study presented in this paper borrows from previous works [4], [5].

A. Definition of Gaussianity

Let T be the timescale of traffic aggregation and let $L_1(T), \ldots, L_n(T)$ be the amount of traffic observed in time periods $1, 2, \ldots, n$ of length T. For any $T > 0$, we want to know if $L(T)$ is Gaussian distributed, i.e., whether $L(T) \sim \text{Norm}(\rho, v(T))$, where ρ is the average traffic throughput and $v(T)$ is the estimated variance of $L(T)$ given by

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

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

B. Analysis of Gaussianity

Quantile-quantile (Q-Q) plots can be used for a qualitative analysis of the Gaussian character of measured traffic. To create a Q-Q plot, the inverse of the normal cumulative distribution function $\text{Norm}(\rho, v(T))$ must be plotted against the ordered statistics of the sampled data $L(t)$. Therefore, the pairs for a Q-Q plot are determined by:

$$\left(\Phi^{-1} \left(\frac{i}{n+1} \right), \alpha_{(i)} \right), \quad i = 1, 2, \ldots, n \quad , \tag{3}$$

where Φ^{-1} is the inverse of the normal cumulative distribution function, $\alpha_{(i)}$ are the ordered traffic averages for each time bin of length T and n the size of our sample (i.e., number of time bins of size T). Note that $\frac{i}{n+1}$ is used instead of $\frac{i}{n}$ because the 100th percentile is infinite for the normal distribution. However, for large sample sizes (i.e., large n), the difference is not significant [11], [12].

Fig. 2 shows Q-Q plots generated from an example trace using two different values of T. For such plots, a traffic sample is considered "perfectly Gaussian" when all the points fall on the diagonal line. By visually analyzing the plots in Fig. 2, one can conclude that, at both T, the traffic from the example trace is "fairly Gaussian", since only few points do deviate from the diagonal line.

When creating Q-Q plots of Internet traffic time series, it is common to see points at the high-end of the plot that fall distant from the diagonal line. This is due to the well known heavy-tail characteristic of traffic. This is a very important characteristic when the context of the study on Gaussianity is related to management tasks such as bandwidth provisioning [6] because such points represent significant fluctuations of traffic that occur at the considered timescale T. In the example of bandwidth provisioning, such fluctuations will impact traffic variance, which is an important parameter for computing the required link capacity for a given input traffic.

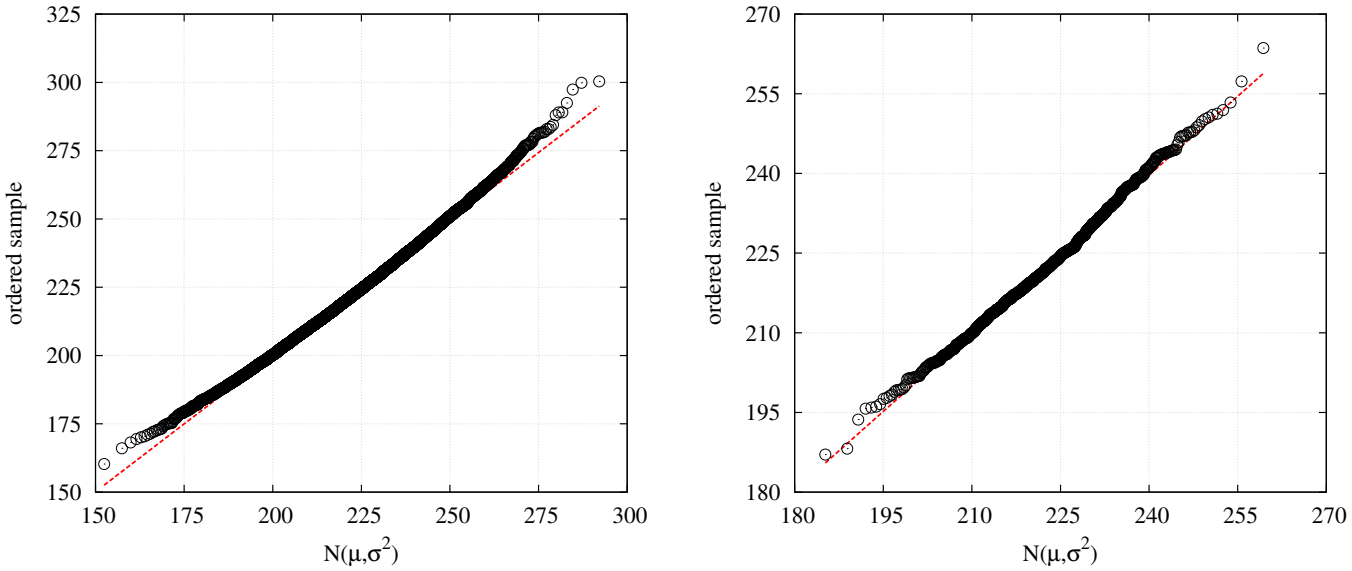
Q-Q plots provide a good visual analysis of the *goodness of fit* of the measured traffic compared to a Gaussian traffic model. However, a quantitative analysis is also needed to support observations from such plots. There are several procedures to quantify Gaussian *goodness of fit*. We opted for the *linear correlation coefficient* [7]. This choice was also done to conform to the methodology followed by previous works [4], [5]. The *linear correlation coefficient* is defined by:

$$\gamma(x, y) = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 \sum_{i=1}^{n} (y_i - \overline{y})^2}} \quad , \tag{4}$$

where the pair (x, y) is the same as in Eq. (3). Clearly, for a given traffic trace, $|\gamma| = 1$ if and only if all points lie perfectly on a straight line in the Q-Q plot. It is important to note that $\gamma \geq 0.9$ corresponds to a Kolmogorov-Smirnov test for normality at significance 0.05, which supports the hypothesis that the underlying distribution is normal. The values of γ for the example trace in Fig. 2 are, respectively, $\gamma_{T=100ms} = 0.9986$ and $\gamma_{T=1s} = 0.9981$.

IV. Results

We have divided our analysis in three parts. First, we assess the impact of *horizontal* traffic aggregation on Gaussianity (Section IV-A). That is, we vary T when analyzing a given input traffic. In a second step, we study the impact of *vertical* traffic aggregation (Section IV-B). Vertical aggregation corresponds to the number of (ideally independent) sources being aggregated. Finally, in



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

Fig. 2. Q-Q plots for a single example trace at different T; this example trace is from location D.

Section IV-C we study the Gaussianity for the long-term measurement from location $F2$.

A. Horizontal traffic aggregation

The goal of this section is to assess whether the Gaussianity *goodness of fit* remains constant over various timescales. That is, we want to find out if a value of γ at one timescale gives indication of how traffic behaves at other values of T. According to [4], larger horizontal aggregation of traffic is needed to justify Gaussian distribution. That is, traffic tends to be more Gaussian-like at larger timescales. The considered timescales for this analysis are based on the same assumption as in [5]: timescales from 5ms to 5s dominate the Quality-of-Service as perceived by the end user. In this work we extend the timescale range to 1ms to 30s.

Each line of the plot in Fig. 3 presents the γ value for one single, randomly selected 15-minute trace per location. Traffic tends to become less Gaussian on shorter timescales mainly when the link aggregate is not too large as, for example, in locations $C1$ and $C2$. That is, the shorter T, the less traffic is aggregated and the higher the variance due to traffic bursts caused by individual sources.

According to [4], a very short T, i.e., in the range of milliseconds, can be too close to the packets' transmission intervals. That would result in a time series with a binary behavior, where we may have or have not packets being transmitted within the period of length T (i.e., ON/OFF behavior). Such a time series is, obviously, not Gaussian. This can be even more problematic if we consider links that aggregate traffic of very few sources, because we may have binary-like traffic time series even in the milliseconds timescale.

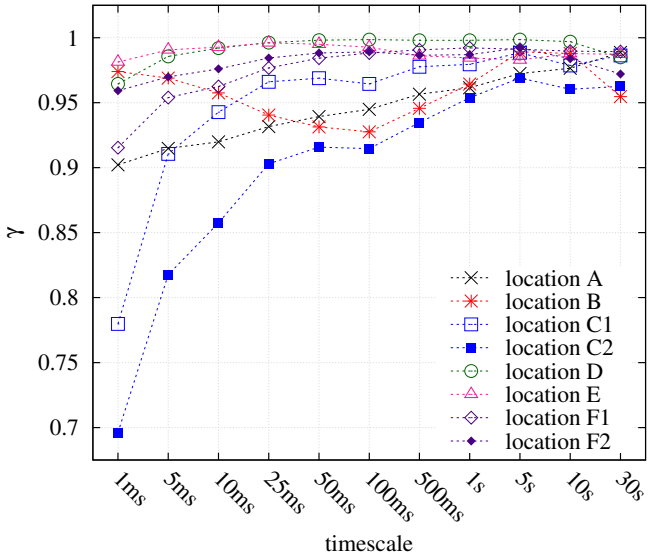


Fig. 3. γ at various timescales for an example trace from each location.

For example, the traces from locations $C1$ and $C2$ show a very bad Gaussianity fit at $T = 1$ms. This problem is alleviated when increasing T over a certain threshold, where γ becomes fairly stable. The same behavior, but with less impact on γ, can be observed for traces from other locations. Interestingly, the example trace from location B in Fig. 3 has a different behavior than others. However, the observed fluctuation of γ through various timescales is not large: γ is larger than 0.9 for all values of T. A similar situation has been also observed in [5].

One important conclusion is that it would not be safe

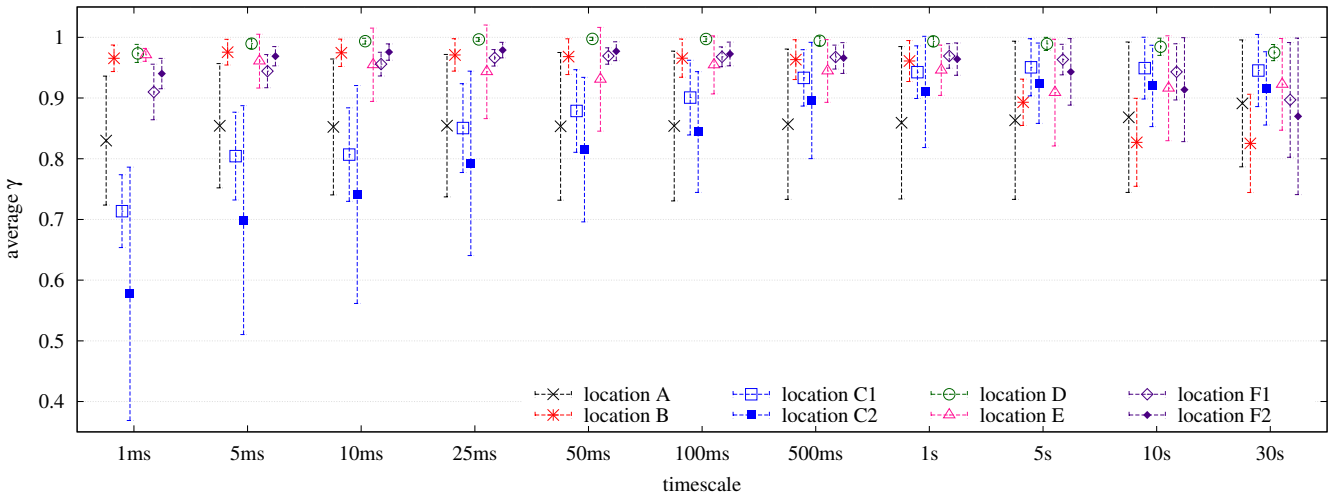


Fig. 4. Average γ at various timescales for all traces in our dataset.

for an arbitrary location to assume Gaussianity at very short timescales. That is, if the traffic is Gaussian at $T = 5$ms, it does not necessarily means that the same traffic will remain Gaussian at $T < 5$ms. Furthermore, Fig. 3 also indicates that γ, to a wide extent, monotonously increases with T, and, hence one can assume Gaussianity at timescale T_1 for a particular traffic if the same traffic is Gaussian at T_0 and T_2 (where $T_0 < T_1 < T_2$).

As an overall representation of the Gaussian goodness of fit of our entire dataset, we have calculated the average γ for all traces. Fig. 4 shows the average γ from all traces for each location at various timescales and the respective standard deviation. The idea here is to find out whether the whole traffic of a location remains Gaussian through different timescales. For all locations, except location B, the average γ increases at higher timescales, or it remains almost constant. Again, this complies with statements from [4] on how Gaussianity should increase with T.

However, as one can see in Fig. 4, location B again behaves differently than other locations. That is, for B at very large timescales, traffic seems to not be Gaussian. Since it is a 24-hour measurement, one possible explanation is that during the overnight period, traffic is very unsteady with some traffic bursts close to each other. For large T, those bursts would be aggregated to a single time bin, resulting in a strongly non-stationary and, hence, non-Gaussian process.

We observe from Fig. 4 that the γ values for the traces of location A and $C2$ depict the largest variation among all locations. In order to allow a better understanding we show the cumulative distribution function (CDF) of γ for all traces per location in Fig. 5, for an arbitrarily chosen timescale of $T = 1s$. For more than 85% of all traces in our dataset $\gamma \geq 0.9$, which means that most traces from our dataset are at least in the "fairly Gaussian" level. For all locations, except A and $C2$, at least 85% of the

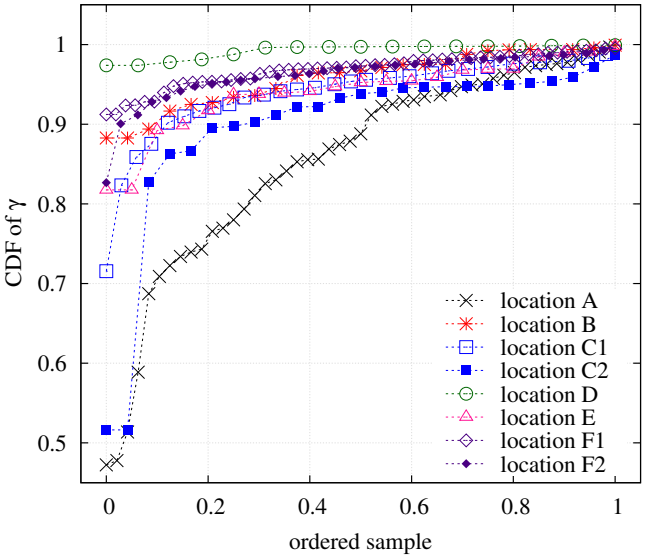


Fig. 5. CDF of γ for all traces per location; $T = 1$s; points sampled for visualization.

location's traces have $\gamma \geq 0.9$. Around 50% of traces from A and 25% from $C2$ have poor Gaussianity fit. Since these two locations are 24-hour measurements from very small networks, the traffic averages significantly decrease during the night due to the reduced number of active hosts, and this causes Gaussianity fit to also decrease. Such problem is further discussed in Section IV-B.

Finally, it is also interesting to know the consistency of γ of a location over all considered timescales. Recall that Fig. 3 shows γ for a set of example traces and Fig. 4 shows the average for all traces from a location at specific timescales. One way to find out the stability of γ for each location is to compute, individually for each

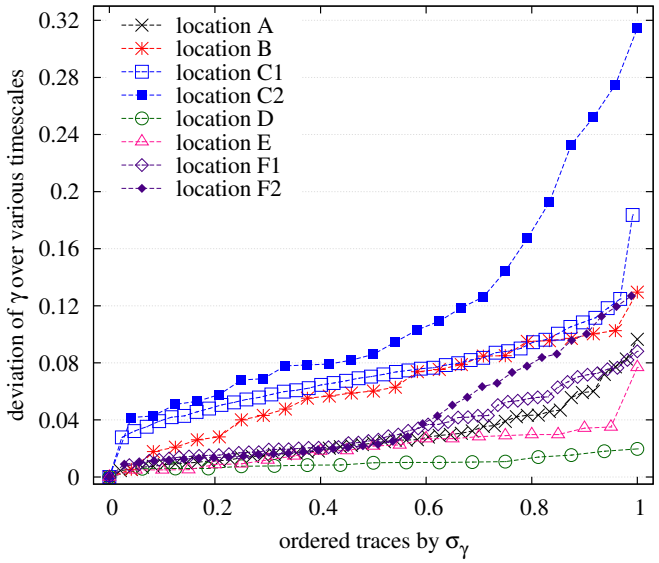


Fig. 6. Variance of γ across various timescales; points sampled for visualization.

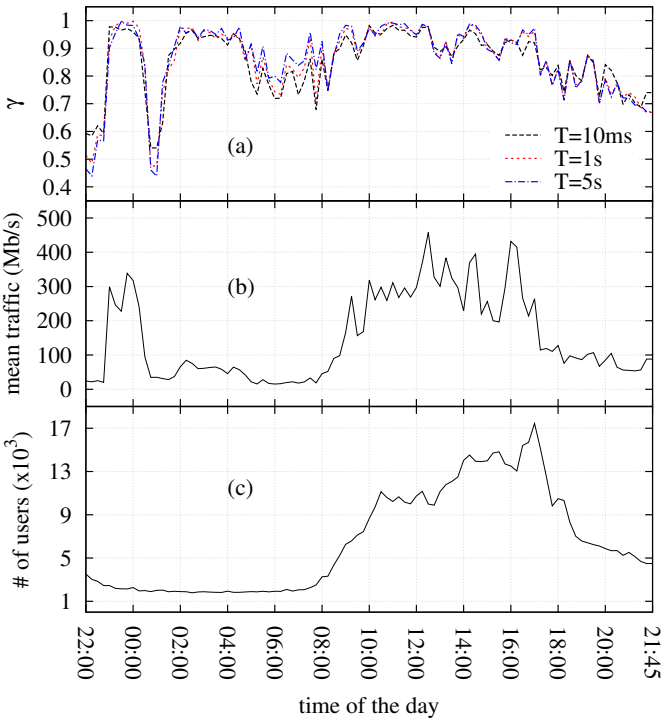


Fig. 7. Gaussianity goodness of fit for a 24-hour measurement period from location A; for b and c, $T = 1$s.

trace, the standard deviation of the trace's γ at various timescales. This metric was proposed in [5]. Hence, for a trace with gamma γ_T for $T = 1ms, \ldots, 30s$, we compute $\sigma_\gamma = \sqrt{\text{Var}[\gamma_T]}$.

Fig. 6 shows the results for each location with the traces sorted by their standard deviation σ_γ. The figure reveals that for more than 50% of all traces $\sigma_\gamma \leq 0.05$, and that for about 90% of all traces $\sigma_\gamma \leq 0.09$. Analyzing each location separately, we see that $\sigma_\gamma < 0.02$ for all traces from D, and for nearly all traces from E, but one, $\sigma_\gamma < 0.04$. For all traces from locations A and $F1$, $\sigma_\gamma < 0.1$. And for the worst cases, the amount of traces that have $\sigma_\gamma \geq 0.1$ for locations B, $C1$, $C2$ and $F2$ are, respectively, 12%, 15%, 42% and 10%. Therefore, traces from these four locations are the ones where γ varies most through different timescales. These results strengthen our previous observations on which Gaussianity is quite constant over timescales, and traffic that exhibits good Gaussian fit at T_0 and T_2 is likely to be Gaussian also at T_1.

B. Vertical traffic aggregation

Previous works [4], [5] have studied the impact of vertical aggregation on the Gaussianity of traffic. Vertical aggregation refers to the amount of aggregated traffic sources. An important question is how many sources are needed to guarantee the Gaussian characteristic of the traffic. Furthermore, one is interested in a definition of traffic source that can be easily used to calculate the number of active sources. For example, a traffic source is not necessarily equivalent to a TCP connection.

In [5], the authors attempt to quantify the number of users (measured as number of observed IP addresses) necessary to justify the Gaussianity assumption. To do so, they sample traffic from randomly selected users and compute γ for it. They conclude that few dozens of users would be enough to justify traffic Gaussianity.

We believe that it can be risky to solely rely on the number of observed users since this assumes that all users behave uniformly in all networks. It is not clear whether the same number of users sufficient for Gaussianity in network X would also be sufficient in network Y. For example, users in a university campus network may behave completely differently from the sources observed in a backbone link. An alternative approach is to relate the level of vertical aggregation to the amount of traffic aggregated. However, this can be also dangerous since individual hosts can also have high transmission speeds, as already observed in [4]. Therefore, we study in the following the impact of vertical aggregation on Gaussianity both in terms of (i) the number of hosts and (ii) the amount of traffic aggregated.

Three 24-hour measurements are used in this analysis. The natural diurnal pattern present in such measurements results in strong variations in the network usage. In addition, the sources that are active during the night often behave quite differently from sources active during daytime. This allows us to study the impact of a wide range of scenarios on the Gaussianity.

For location A, all 96 traces are used, that is, the whole traffic in the 24-hour period is considered. Fig. 7a shows γ for each 15-minute trace. For all T, one can see that

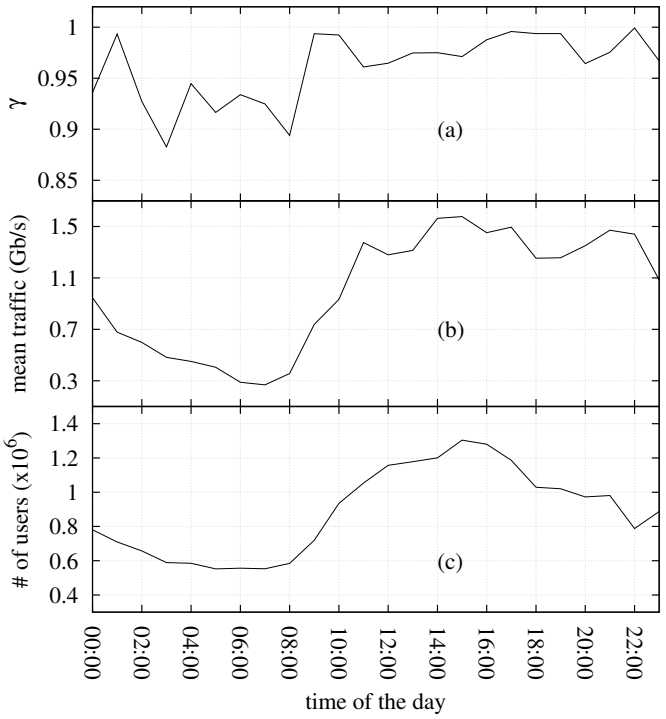


Fig. 8. Gaussianity goodness of fit for a 24-hour measurement period from location B; $T = 1$s.

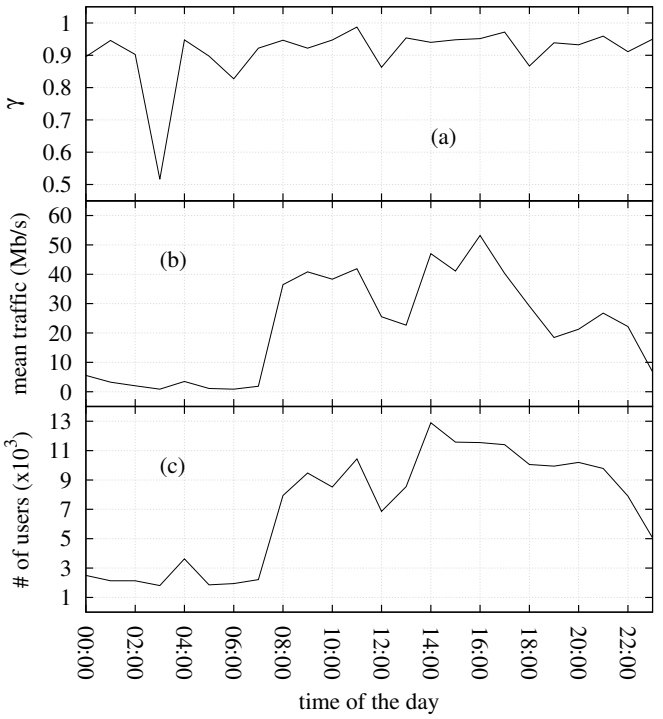


Fig. 9. Gaussianity goodness of fit for a 24-hour measurement period from location $C2$; $T = 1$s.

γ oscillates a lot across the measurement period. Good Gaussianity fit is found even during the overnight period, what would not be expected considering the small number of active users in the network (shown in Fig. 7c). However, Fig. 7b shows that between 23:00 and 01:00, there is an increase on the traffic average which seems to be the reason of the good fit. A smaller increase can also be observed from 2:00 to 4:00. This might have been the result of automatic operations, such as overnight backups. The figure also shows that traffic averages during the day are generally higher due to the higher number of active users in the network. Consequently, Gaussianity of traffic tends to be more regular in the busiest period of the monitored link.

The results indicate that Gaussianity depends more on the behavior than on the quantity of active users. Although we have more than thousand active hosts at 1:00 and 21:45, the Gaussianity fit is low. Furthermore, Fig. 7 shows that a high traffic rate can be a better indicator for good Gaussianity than the number of users. Note that the opposite is not necessarily true.

We have also studied the Gaussianity fit of a 24-hour measurement from location B. However, in this measurements we have only the first 15 minutes of each hour during an entire day. The results of this analysis are presented in Fig. 8 for $T = 1$s.. The measured link in location B has many more active hosts than location A, also during the night. One of the main reasons is that this link also transports traffic from the residential buildings located on the university campus and the public servers.

There are only two moments in which Gaussianity fit of traffic is not good enough, i.e., $\gamma < 0.9$: at 03:00 and at 08:00. In these moments, traffic averages and number of users were quite low compared to other periods of the day. Although one cannot argue that Gaussianity is as bad as observed for the overnight period in Fig. 7, it is clear that it is unstable between 00:00 and 09:00. Again it seems that the behavior of a few users determine the Gaussian characteristic during the light-loaded period of the link. Again, we observe that a high traffic rate is a better indicator for a high γ value than the number of users: For example, γ closely follows the traffic rate in the period from 19:00 to 23:00. Again, the opposite is not necessarily true.

Finally, we have the same kind of measurement for location $C2$, where only the first 15 minutes of traffic of every full hour was measured during an entire day. One important remark regarding location $C2$ measurements is that, unlike $C1$, the measurement took place when most students were on holidays. That was unfortunate since we missed a portion of network users, but at the same time it gave us the possibility to study the properties of traffic aggregated from fewer users. Another remark is that this university does not have residential buildings on the campus and, hence, overnight traffic averages are likely to be low. Fig. 9 presents the results for this location.

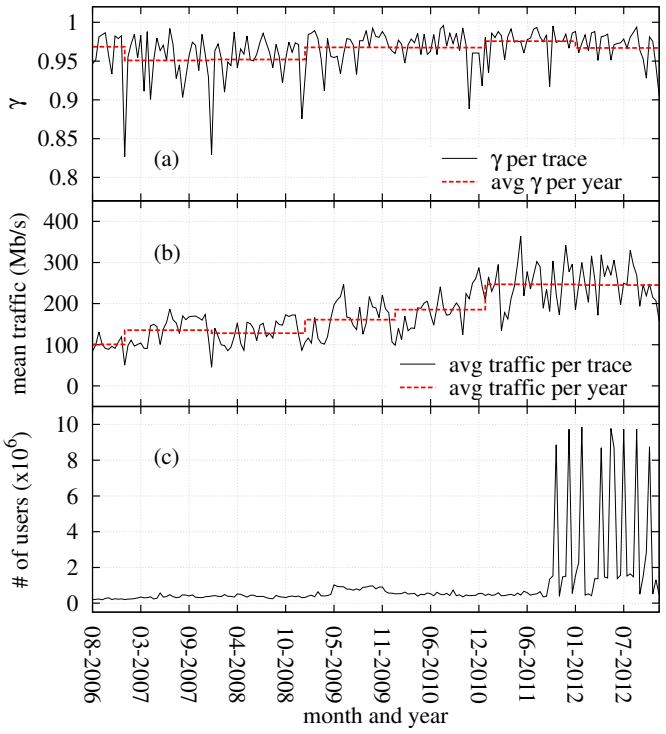


Fig. 10. Gaussianity goodness of fit for traces from location $F2$; $T = 1$s.

As one can see from Fig. 9a, γ is not very high (close to 0.9) but fairly stable. Indeed, it seems that it is affected neither by the traffic rate nor by the number of users. We believe that, due to the holiday period, the traffic characteristics mainly arise from the rather constant behavior of the employees and automated processes, while variations caused by, for example, file transfers are rare. The main take away from Fig. 9 is that with links that have low capacity or low activity the overall user behavior becomes the dominating factor.

C. Impact of long-term traffic evolution

The aim of this analysis is to check whether long-term traffic evolution has an effect on Gaussian characteristics of traffic. In the context of this paper, traffic evolution would be caused by applications that emerged in the past years. On one hand, services such as Facebook would be responsible for many connections with few transferred data (i.e., short flows) [13]. On the other hand, online video streaming and cloud storage services would be responsible for connections with, generally, large amount of transferred data [13], [14].

Fig. 10 shows the Gaussian goodness of fit γ calculated for the traces from location $F2$ dating from 2006 to 2012. A total of 178 traces (about 2 to 3 traces per month from August 2006 to December 2012) were used in this analysis. We observe that the (already good) average goodness of fit has only slightly increased from 2006 to 2012, while traffic bandwidth has more than doubled. It should be noted that the increased bandwidth results in less variation of γ.

The number of users has not changed significantly during the measurement period. For unknown reasons, the measured link experienced several moments of huge peaks on number of hosts transferring data from September 2011 to December 2012. However, this did not result in higher traffic rates or a better Gaussianity fit.

V. Conclusions

The assumption of Gaussian traffic is widely used in network modeling. However, the most recent systematic study on the presence of Gaussianity in real network traffic is from 2006, relying on measurements from 2004. In this paper, we have verified the assumption of Gaussianity on recent traffic measurments. Our dataset comprises extensive measurements from four continents and covers diverse scenarios, from small campus networks to 10 Gb/s backbone links.

Our results show that the assumption of Gaussianity still holds for current network traffic, indicating that the evolution of the Internet in the past years has not had a significant impact on its Gaussian characteristic. Indeed, most of the analyzed measurement locations show a high or very high degree of Gaussianity for a wide range of considered aggregation timescale. However, this degree can vary depending on the level of vertical aggregation and is usually highest during the busiest period of the network, i.e., during daytime.

Our findings also suggest that it is safer to relate the degree of Gaussianity to traffic bandwidth than to the number of users for high-speed links. The number of active users is less reliable as indicator for Gaussianity because users from different networks may behave differently. Cases in which traffic is Gaussian even with few users and low traffic averages require a deep manual study of the traffic properties, and this is planned as future work.

Finally, we have illustrated the invariance of the Gaussianity property by our study of a trans-Pacific backbone link over a period of six years. Although the amount of traffic transported by that link has considerably changed during the measurement period, the degree of Gaussianity has nearly stayed constant. To the best of our knowledge, this is the first time such a longitudinal analysis has been performed.

We conclude that mathematical models relying on the Gaussianity of network traffic are still valid, especially for the, from the viewpoint of network operators, most interesting periods of high network activity.

Acknowledgements

This work has been supported by the FP7 Univer-Self Collaborative Project (#257513), and by the FP7 FLAMINGO Network of Excellence Project (CNECT-ICT-318488). We thank Idilio Drago for his help on setting up measurements used in this paper.

References

[1] W.E. Leland and M.S. Taqqu and W. Willinger and D.V. Wilson, *On the Self-Similar Nature of Ethernet Traffic*, IEEE/ACM Transactions on Networking, vol. 2, issue 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, issue 3, pp. 226–244, 1995.

[3] I. Norros, *A storage model with self-similar input*, Queueing Systems, vol. 16, issue 3–4, pp. 387–396, 1994.

[4] J. Kilpi and I. Norros, *Testing the Gaussian approximation of aggregate traffic*, in proc. of the 2nd ACM SIGCOMM Internet Measurement Workshop (IMW), pp. 49–61, 2002.

[5] R. van de Meent, M. Mandjes and A. Pras, *Gaussian Traffic Everywhere?*, in proc. of the IEEE International Conference in Communications (ICC), vol. 2, pp. 573–578, 2006.

[6] A. Pras, L. J. M. Nieuwenhuis, R. van de Meent and M. R. H. Mandjes, *Dimensioning Network Links: A New Look at Equivalent Bandwidth*, IEEE Network, vol. 23, issue 2, pp. 5–10, 2009.

[7] B. M. Brown and T. P. Hettmansperger, *Normal Scores, Normal Plots and Tests for Normality*, Journal of the American Statistical Association, 91(436), pp. 1668–1675, 1996.

[8] The CAIDA UCSD Anonymized Internet Traces 2011 - *2011-05-19, 2011-07-21, 2011-12-22*. Available at http://www.caida.org/data/passive/passive_2011_dataset.xml

[9] The CAIDA UCSD Anonymized Internet Traces 2012 - *2012-01-19, 2012-02-16*. Available at http://www.caida.org/data/passive/passive_2012_dataset.xml

[10] MAWI Working Group Traffic Archive. Available at: http://mawi.wide.ad.jp. Last access on December 2012.

[11] L. Makkonen, *Bringing Closure to the Plotting Position Controversy*, Communications in Statistics – Theory and Methods, vol. 37, issue 3, pp. 460–467, 2008.

[12] L. Makkonen, M. Pajari and M. Tikanmäki, *Closure to "Problems in the extreme values analysis"*, Structural Safety, vol. 40, issue 1, pp. 65–67, 2013.

[13] V. Gehlen, A. Finamore, M. Mellia and M. M. Munafò, *Uncovering the Big Players of the Web*, in proc. of the 4th International Workshop on Traffic Monitoring and Analysis (TMA), ISBN: 978-3-642-28553-2, pp. 15–28, 2012.

[14] I. Drago, M. Mellia, M. M. Munafò, A. Sperotto, R. Sadre and A. Pras, *Inside Dropbox: Understanding Personal Cloud Storage Services*, in proc. of the ACM Internet Measurement Conference (IMC), pp. 481–494, 2012.

[15] J. L. García-Dorado, A. Finamore, M. Mellia, M. Meo and M. Munafò, *Characterization of ISP Traffic: Trends, User Habits, and Access Technology Impact*, IEEE Transactions on Network and Service Management, vol. 9, issue 2, pp. 142–155, 2012.