Testing the Gaussian approximation of aggregate traffic

Jorma Kilpi, Ilkka Norros

Abstract— **We search for methods or tools to detect whether the 1-dimensional marginal distribution of traffic increments of aggregate TCP-traffic satisfy the hypothesis of approximate normality. Gaussian approximation requires a high level of aggregation in both "vertical" (source aggregation) and "horizontal" (time scale) directions. We discuss these different concepts of aggregation first separately, with an example from real data traffic, and show how to rule out cases where the level of aggregation will not be sufficient. Gaussian approximation is then quantified with the square of the linear correlation coefficient in normal-quantile plots. We propose an elementary method based on this correlation test, by looking at the behavior of the test statistic for different sample sizes, and show positive and negative examples from the example data. We use this method to look for the first time scale, where the Gaussian approximation is plausible with the example data, and then we look how much more vertical aggregation would be needed for smaller time scales in order to obtain a reasonable approximation by normal distribution.**

Keywords— **Traffic Modeling and Measurements, Aggregate traffic, TCP/IP, Quantile-Quantile plots, Normal plots, Correlation tests.**

I. Introduction

GAUSSIAN modeling of teletraffic became an interesting possibility when the extreme complexity and long-range dependence of computer traffic were revealed in the famous Bellcore measurements [1]. The structurally simplest and best understood long-range dependent stochastic process is the self-similar Gaussian one, fractional Brownian motion (fBm). Therefore, it was used as queue input model in [2], [3]. Because of the simplicity of the model, it has become rather popular in Internet traffic modeling. It is, however, definitely wrong to interpret it as "the" Internet traffic model (remember that the Poisson process really is with good grounds *the* basic model of telephone call arrival process), and the reason is its Gaussian nature. Second-order self-similarity is an informative characteristic only if the traffic is Gaussian — as many examples show, traffic traces with same second-order structure can be totally different as regards, for example, queueing behavior. Although it seems that long-range dependence, i.e. approximate second-order self-similarity is an inherent feature of most Internet traffic, Gaussian character requires additionally a high level of *traffic aggregation* that is often not present. Usually the distribution of Internet traffic in a small time interval has a strongly asymmetric distribution with a rather heavy tail.

J. Kilpi and I.Norros are with the VTT Information Technology, Finland. E-mail: {Jorma.Kilpi,Ilkka.Norros}@vtt.fi.

On the other hand, once the aggregation level is sufficient for using a Gaussian approximation, one can relax the exact self-similarity assumption and consider other Gaussian models than fBm. The Central Limit Theorem (CLT) serves as a general motivation for such models. As regards queueing theory, it turns out that although exact results are non-existent with non-Brownian input, the usual large deviation estimates (first proven for long-range dependent traffic in [4]) seem to be rather accurate, even for small buffers. This technique has been presented and tested with Gaussian simulations for simple queues in [5], [6] and for priority queues in [7], [8]. A short review of COST257 work on Gaussian traffic and queueing models is contained in the COST257 Final Report [9].

Why should the traffic be Gaussian? Note that CLT alone is never sufficient: at the time scale of a minimum size packet, the traffic is always on/off: either there is a transmission going on at link speed, or there is silence. Both *vertical* aggregation, i.e., presence of a large number of independent traffic flows, and *horizontal* aggregation, i.e., working at a sufficiently large time scale, are needed to justify Gaussian modeling. Theoretically, vertical aggregation can be expected to work through CLT and horizontal aggregation through generalizations of Donsker's theorem (see, e.g., [10]). However, even these cannot be relied on *a priori*, since it may well happen that the aggregate converges to a Gaussian sequence too slowly to make sense in traffic modeling, or does not converge at all (for example, the proper horizontal limit process could be a Lévy process; see [11] or [12]).

In this paper we search for methods or tools to decide,

what are the necessary and sufficient levels of aggregation which make the traffic Gaussian, where these levels approximately lie, and how to test the normal distribution assumption for strongly dependent time series of aggregate TCP/IP traffic. We emphasize two aspects in this context. First, what we would actually like to do is to understand the minimal level of an approximation that still makes sense from traffic theory point of view, i.e. when estimating the probability of buffer overflow or the probability of link speed exceedance. Secondly, as related to the first aspect, we are hence not looking for best-fit model in the strict statistical sense, which might by accident be some other distribution, since CLT alone does not lead to for example log-normal, Gamma or Weibull distributions. However, we will use log-normal distribution as an example of positively *skewed* alternative to normal distribution. This is natural, since the same methods can be applied to test both normal and log-normal approximations.

The criterion of Gaussian approximation is based on well-known normal-quantile (N-Q) plots and related correlation tests that compare the empirical distribution with a fitted Gaussian distribution, see e.g. [13], [14] or [15]. The main problem in our context is that the usual independence assumption does not hold in time direction and the independence of traffic sources can be assumed only if the capacity is not a restriction.

To show examples of the methods, we use a special type of traffic that can be expected to be perhaps the most suitable for Gaussian modeling. Our data comes from `tcpdump` traces captured from home users' Internet TCP/IP traffic in the access network of a large ISP. In this data set, the customers were connected over ISDN and telephone lines. The maximal user speed is thus mostly 64 kb/s, sometimes 128 kb/s. We will study the data through different resolutions, starting from 1 milliseconds (ms) up to four seconds (4096 ms). Because of the low user speed, there is a definite upper bound of the effect that a single user can contribute in a time slot of fixed resolution to the aggregate traffic. So, if the level of aggregation (both vertical and horizontal) is high enough such that individual sources are swallowed, then, due to the CLT, Gaussian approximation should work well for aggregated home user traffic. We shall show examples of positive and negative cases, and give some explanations for the negative cases.

Our methods are not restricted to the example data set, but in order to understand the conclusions why the Gaussian approximation works or does not work, we will start by describing the data set in section II. Another partially data-specific section III studies the simple necessary (but unsufficient) criterion that empty time slots must be rare for good Gaussian approximation. In section IV we search for a method to measure the goodness-of-fit to normal distribution in the case when the original sample does not consist of independent observations. In section V, we show evidence of whether the assumption of stationary time series is violated or not, look for the first time scale where the Gaussian approximation is plausible and try to estimate how much more vertical aggregation would be needed in smaller time scales to obtain good Gaussian approximation. The conclusions are drawn in section VI.

II. Description of the data set used in the examples

All real data examples used in this paper are taken from a single data set, consisting of `tcpdump` records of TCP packets from a measurement point of a dial-up Internet connection service[1]. Dial-up users were identified using additional information from a simultaneous authentication log file.

This data set was deliberately chosen from a much larger measurement (10 days) for the purposes of this paper and thus should be considered as a reference data. While it represents typical traffic from its measurement point, we cannot infer that its properties are general but it serves as an example of the ideas and methods that we will present.

The traffic trace was recorded on Thursday, March 2, 2000, at 15:10:50-17:42:30. Figure 1 shows that especially for the last 30 minutes, the traffic looks very stationary (as well as Gaussian, and long-range dependent (LRD)). The clear increase of the traffic during the first minutes after 17:00 must be considered as a non-stationary, i.e. a deterministic feature, because 17:00 was a daily time of tariff change. In order to remove the non-stationary effects we restrict the examples mostly to the last 30 minutes of the trace, and give some comments from first 90 minute period. The figure 1 shows a clear increase in vertical aggregation for the last 30 minute period.

Table I presents the basic information about our traffic trace. "Upstream" means from the user to the network. The item "Active sessions" is the number of dial-up sessions that contributed to TCP-traffic by at least one packet. There were 1535 sessions that contributed in both directions.

If the users must compete for the same finite resources, they cannot be considered independent. The TCP/IP-traffic rate at this measurement point generated by dial-up users can be significantly higher than in the present data, hence there are no signs that the system would be even close to its performance limits, and for this reason we assume that the individual user streams behave inde-

[1]Call level analysis of similar data was presented in [16].

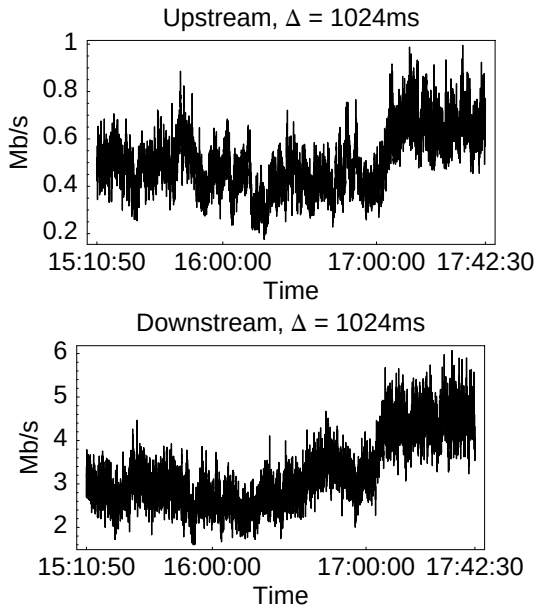

Fig. 1. Time series of up- and downstream TCP traffic with resolution of 1.024 seconds. Note that the scales of the vertical axes of the two pictures are different.

TABLE I
GLOBAL CHARACTERISTICS OF THE TRAFFIC TRACE.

Direction	Upstream	Downstream
Duration (s)	9 100	9 100
Volume (B)	573 167 180	3 738 254 539
# of packets	5 635 280	6 978 990
Active sessions	1 557	1 537

pendently.

The activity of the users varies a lot. Figure 2 shows volumes per session in both up- and downstream directions. In that figure sessions are ordered in the x-axes by their downstream volume, the continuous line showing this. The point in the cloud with the same x-coordinate describes the corresponding upstream volume.

Figure 3 shows the packet size distributions in both directions. Typical packet sizes in both directions were 40, 576 and 1500 bytes, the other sizes being very rare. In the upstream direction over 80% of packets were of size 40 bytes and also in the downstream direction almost 20% were of that size. This means that for small time resolutions there were a lot of time slots with just one or two packets carrying no data with them and a lot of empty time slots. It is clear that then a continuous distribution function cannot be even an approximate model.

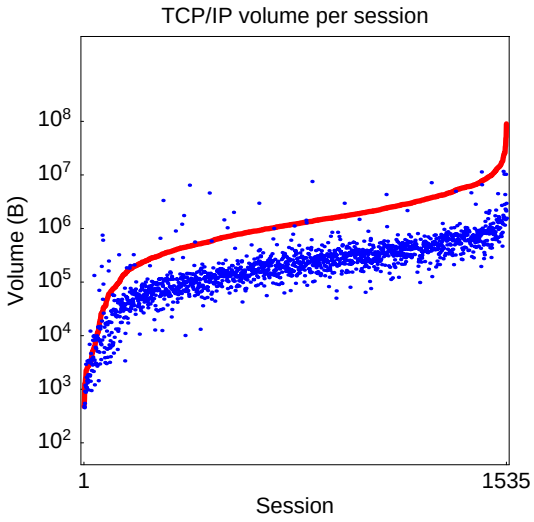

Fig. 2. Up- and downstream volume per session.

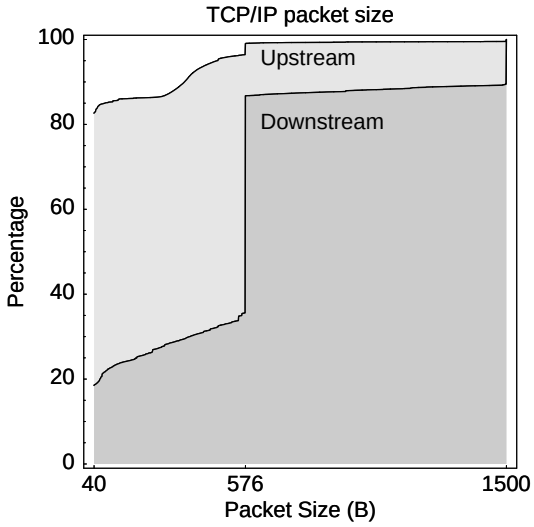

Fig. 3. Cumulative packet size distributions.

III. AT WHAT RESOLUTION ARE EMPTY TIME SLOTS EXCEPTIONAL?

An elementary necessary criterion for a non-negative time series looking like Gaussian is that zeroes should be rare. Let us look for rough a priori arguments for guessing a time resolution below which we cannot even hope for a good Gaussian fit for the example data because of too many empty time slots.

A. Horizontal aggregation

For resolutions larger than 1 ms a full size packet of 1500 bytes can be considered as a point-like entity in the 100 Mb/s Ethernet measurement interface.

We define the *horizontal minimum rate* for resolution Δ by $M(\Delta) = (\text{ mean packet size })/\Delta$. The observed mean

packet sizes were 102 bytes upstream and 536 bytes downstream. (The exact value of $M(\Delta)$ is not crucial though).

The idea of horizontal minimum rate is that when the aggregate traffic has a mean rate above $M(\Delta)$, then "typical" time slots of length Δ contain at least one "mean size" packet. Especially, this rules out large frequency peaks of traffic rates caused by 0, 40 or 80 bytes per single time slot in the corresponding histogram of traffic rates, making a continuous distribution model a little bit more reasonable one. Figures 4 and 5 show this area for up- and downstream directions.

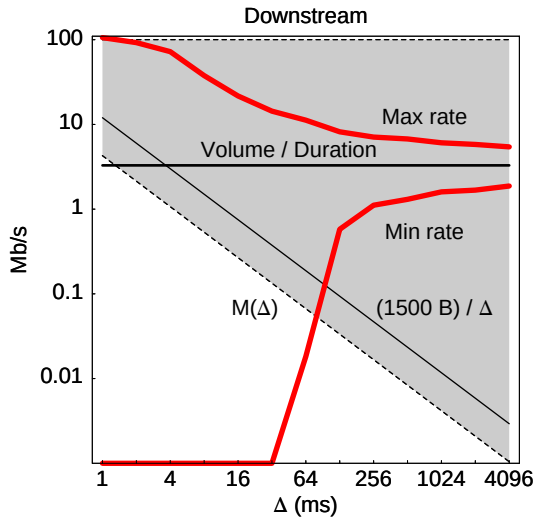

Fig. 4. Shaded area shows when the traffic rate exceeds $M(\Delta)$. Note that both axes are logarithmic.

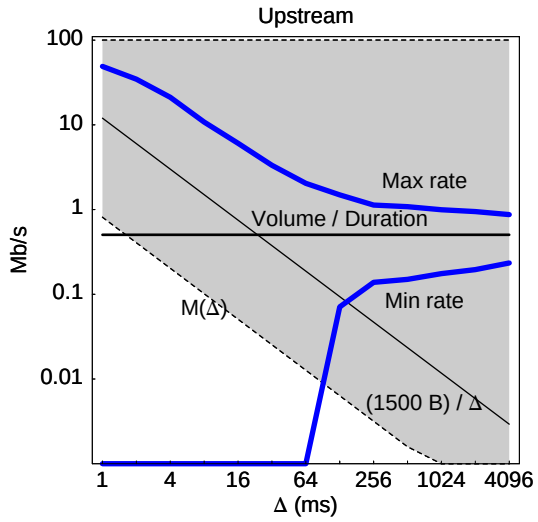

Fig. 5. The upstream case.

The horizontal solid line in figures 4 and 5 presents the total number of bytes in the corresponding direction observed during the whole measurement duration divided by the duration. The decreasing solid line shows the rate obtained by a single 1500 bytes packet per Δ. The curves present, for each resolution Δ, the empirical maximum and minimum amounts of bytes observed in our data in time slots of length Δ, divided by Δ.

B. *Vertical aggregation*

Let us then keep the resolution Δ fixed and look for a number of sources that is sufficiently high to keep the traffic non-negative in most time slots. We use the following strongly simplified model.

Assume that the sessions start according to a stationary Poisson process, their lengths are independent and identically distributed (i.i.d.), and each session i generates traffic with constant random rate X_i, with unit kb/s. The X_i's are i.i.d. as well. We assume finite mean and variance: $\mathrm{E}X_j = \mu$ and $\mathrm{Var}(X_j) = \sigma^2 > 0$.

The random variables X_j are estimated from data by the numbers x_j, where the value of x_j represents all the bytes or bits coming from the $j:th$ session divided by the duration between the first and the last packet. The sessions are numbered (ordered) by the first observations of their TCP-contribution. Thus the numbering may differ for up- and downstream directions.

The number of contributors in each subinterval varies. We would like to know how many contributors we should at least have in order to keep the aggregate traffic rate for any subinterval of length Δ mostly above $M(\Delta)$. More precisely, let us fix an exceeding probability $0 < p < 1$ and find the smallest number n such that

$$\mathrm{P}\left\{\sum_{j=1}^{n} X_j > M(\Delta)\right\} \geq p. \qquad (1)$$

A normal approximation yields

$$\mathrm{P}\left\{\sum_{j=1}^{n} X_j > M(\Delta)\right\} \approx 1 - \Phi\left(\frac{M(\Delta) - \mu n}{\sigma\sqrt{n}}\right),$$

and we choose as our candidate for n the minimal solution of

$$\frac{M(\Delta) - \mu n}{\sigma\sqrt{n}} \leq \Phi^{-1}(1-p).$$

Let $p = 0.99$. The edge of the shaded area of figure 6 shows the minimal n as a function of the resolution Δ. The curves show the empirical maximum and minimum numbers of contributors in a time slot of each resolution. The max curve crosses the edge between 32 and 64 ms. We deduce that, for these data, reasonable Gaussian approximation cannot be expected at all with smaller resolutions than 32 ms.

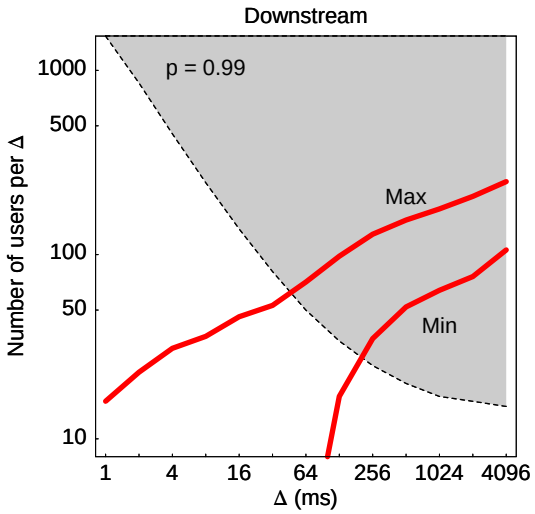

Fig. 6. The empirical maximum and minimum number of contributors for each resolution, and the shaded region of suffiently large values of n, related to the exceeding probability $p = 0.99$. Please note again that both axes are logarithmic.

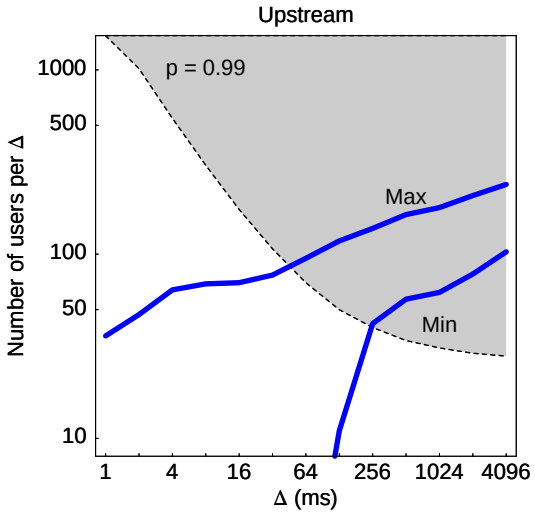

Fig. 7. The upstream case.

C. Conclusions

We derived two approximate criteria for ruling out cases where it is not even worth of making N-Q plots for checking the quality of a Gaussian approximation. For CLT to work, we should be all the time inside the shaded area of figure 6 (or 7 in the upstream case). Note that we are then practically automatically inside the shaded area of figure 4 (5). This effects on resolutions 64 ms and 128 ms only. Of course, the true threshold curve for plausible Gaussian approximation turns out to be considerably higher than those obtained by our simple necessary conditions, but these considerations already gave us some idea. We have obtained a rough curve, above which we can hope the CLT will do its work.

IV. Description of the methods

For a fixed resolution Δ we are observing a time series of bytes per $i:th$ time slot of length Δ, denoted by x_i. These values x_i are assumed to have a marginal distribution F. We would like to formulate our null hypothesis as

$$H_0 : \ F \approx \Phi_{\mu,\sigma}, \tag{2}$$

where $\Phi_{\mu,\sigma}$ is a cdf of some $N(\mu,\sigma)$ distribution. The main problem we have with this hypothesis (2) above is that we have no tools yet to even define what the "$\approx$" would mean. For this reason we formulate the null (composite) hypothesis traditionally as

$$H_0 : \ F = \Phi_{\mu,\sigma}, \tag{3}$$

for then we have some traditional methods to use. The main difficulty now is that the observations x_i are not independent but indeed rather strongly correlated.

We used well-known *quantile-quantile* (Q-Q) and especially *normal quantile* (N-Q) plots for testing the Gaussian approximation. A N-Q plot presents the pairs

$$\left\{ a_i, x_{(i)} \right\}, \quad i = 1, \ldots, n, \tag{4}$$

where n is the sample size, $x_{(1)} < \ldots < x_{(n)}$ are *the order statistics*, i.e. increasingly ordered observations, and the points a_i are *the plotting positions* (see [13] and the references from there). The plotting positions are always assumed to be ordered, $a_1 < \ldots < a_n$ and, since the normal distribution is symmetric, a natural additional property to require will be that $\sum_{i=1}^{n} a_i = 0$. We will describe two natural choices for the plotting position vector $a = (a_1, \ldots, a_n)$ in a section IV-B below.

A. Two-sample test

We describe shortly a well-known (non-parametric) Q-Q plot method which is used to test whether two samples are taken from the same distribution. Recall that the empirical cdf

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{]-\infty,x]}(x_i)$$

has a well defined generalized inverse

$$F_n^{-1}(t) = \inf\{ \ x \in R \mid F_n(x) \geq t \ \},$$

and if $\frac{i-1}{n} < t \leq \frac{i}{n}$, then $F_n^{-1}(t) = x_{(i)}$. If we have two disjoint samples, x_i and y_i, $i = 1, \ldots, n$, and we make a plot

$$\left\{ x_{(i)}, y_{(i)} \right\}, \quad i = 1, \ldots, n, \tag{5}$$

then a linear shape in the plot suggests that the quantiles of the two empirical distributions coincide, i.e. samples are taken from the same underlying distribution F.

In the case of a time series, with nearby samples, a non-linear shape in the plot can be used to show that the assumption of a stationary time series does not hold. We will consider a linear shape as an evidence for that the assumption of a stationary time series is justified.

B. Choice of plotting positions

If $1 \leq i \leq n$, then $\frac{i-1}{n} < \frac{i}{n+1} < \frac{i}{n}$ and the first natural choice for a is to define

$$a_i = \Phi^{-1}\left(\tfrac{i}{n+1}\right), \quad i = 1, \ldots, n, \tag{6}$$

where Φ is the cdf the standard normal distribution. In this case we have by the symmetry of the normal distribution that $\sum_{i=1}^n a_i = 0$. The use of Φ instead of some $\Phi_{\mu,\sigma}$, where μ and σ are known or estimated beforehand is justified since $\Phi_{\mu,\sigma}^{-1} = \mu + \sigma\Phi^{-1}$ and a linear change of plotting positions should not affect the linear shape in the plot. Then the plot (4) compares the quantiles of the standard normal distribution with the quantiles of the empirical distribution.

The second choice for a was introduced in [13]. We start first by minimizing the expression (see [13] for more details)

$$\int_0^1 \left(F_n^{-1}(t) - \mu - \sigma\Phi^{-1}(t) \right)^2 dt \tag{7}$$

with respect to μ and σ. Straightforward calculations give $\widehat{\mu} = \overline{x}$ and

$$\widehat{\sigma} = \int_0^1 F_n^{-1}(t)\Phi^{-1}(t)\, dt = \sum_{i=1}^n b_i x_i,$$

where

$$\begin{aligned} b_i &= \int_{(i-1)/n}^{i/n} \Phi^{-1}(t)\, dt \\ &= \phi\left(\Phi^{-1}(\tfrac{i-1}{n})\right) - \phi\left(\Phi^{-1}(\tfrac{i}{n})\right). \end{aligned}$$

Using the notations $\phi_i = \phi\left(\Phi^{-1}(\frac{i}{n})\right)$ with $\phi_0 = \phi_n = 0$, the plotting positions are defined by

$$a_i = \frac{\phi_{i-1} - \phi_i}{\sum_{i=1}^n (\phi_{i-1} - \phi_i)^2} \quad i = 1, \ldots, n. \tag{8}$$

Here ϕ is the pdf of the $N(0,1)$ distribution. In this case $\sum_{i=1}^n a_i = 0$ since this is a telescoping sum. These plotting positions (8) extend further into the tails than (6), otherwise they do not differ much. We chose to use (8) and call them "smooth" quantiles.

The reason for choosing these smooth quantiles is that the square root of the integral (7) with $\mu = \overline{x}$ and $\sigma = \widehat{\sigma}$ is a distance between the empirical cdf F_n and the family of (univariate) normal distributions in a particular metric, see [17] and [15] for further references in that direction. In the papers [17] and [15], the metric is called *the Wasserstein metric*. However, we do not know yet whether we could use this metric explicitly in this context.

A good fit to the Gaussian distribution means that all points of the N-Q plot are close to the diagonal. If we first take logarithms of data and after that make N-Q plots, we get plots which can be used for testing the log-normal approximation of a sample.

C. How to measure the goodness of fit?

Originally the N-Q plot was a visual tool to detect non-normality from small samples. Our usage is completely the opposite: we want to estimate the goodness of fit of a normal approximation for very large samples. The large sample size creates a visual problem of which type of deviations from linear shape are significant. Figure 8 shows an example of a N-Q plot where the sample size is 1760. The linear shape in the main body looks extremely good, the tails deviate little but clearly.[2] The situation is not always as good as the figure 8 might suggest. When enlarging the sample size, a clear linear shape can sometimes become much worse, and the linear shape can reappear if the sample size is again enlargened. The deviations at the (upper) tail may look worse and then less worse. Also, if two different models for the same sample (like normal and log-normal in our case) are both quite close but neither of them is definitely good, it can be hard to say which one deviates more from the linear shape. So some quantitative criteria have to be used.

As a quantitative measure of linearity, or *goodness of fit*, we used r^2, the square of *the linear correlation coefficient* r (see e.g. [14] or [13]), defined generally for arbitrary, not necessarily ordered, vectors $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ by

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

[2] Large sample size may create also other problems: a statistical test may fail due to too large sample, or numerical accuracy might become a problem.

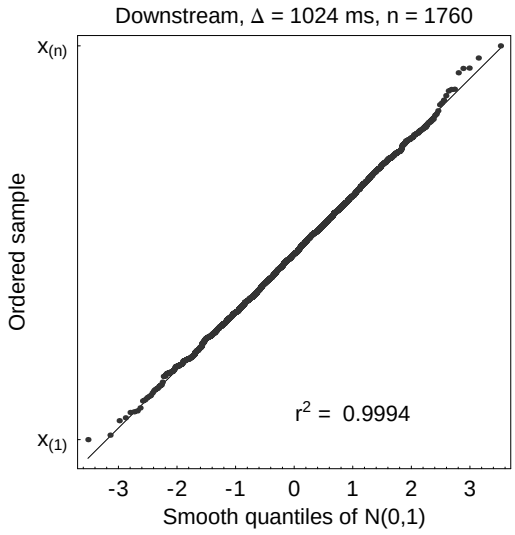

Fig. 8. One of the best data examples of N-Q plot together with a minimum least squares sense fitted line. The intercept and slope of the line give such estimates of μ and σ, namely $\overline{x}$ and $\widehat{\sigma}$, that the normal distribution with these estimates as parameters is the closest of all normal distributions to the empirical distribution F_n, in the sense of the Wasserstein metric.

where $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Clearly, $|r| \leq 1$ and $r = 1$ if and only if $y = \lambda x$ for some real number λ, i.e. all the points lie perfectly on a straight line. The value of r^2, and also r, is invariant under location and scale changes, which makes it appropriate for testing the composite hypothesis of normality. It is also intuitively appealing. Only the value of the numerator depends on the order of values x_i or y_i. The distribution of r^2 (whether the null hypothesis holds or not) depends on the sample size n, which is emphasized by writing r_n^2 instead of plain r^2.

The main problem of using this *correlation test* is the absence of p-values — they are practically calculable only in the case when the original observations are independent (see [14] or [15]). If this independence assumption does not hold, the convergence of any test statistic to the limiting distribution may even fail totally, or at least be essentially slower than in the independent case, see e.g. chapter 10 of Beran's book [18].

For example, in the *Shapiro-Francia* test (a large sample approximation of the *Shapiro-Wilk* test, see [19] and [20])

$$a_i = EZ_{(i)}, \quad i = 1, \ldots, n, \tag{9}$$

the expected values of random variables Z_i, where Z_i, $i = 1, \ldots, n$, is a random sample of *independent* observations from a standard normal distribution. In this case $\frac{i-1}{n} \leq \Phi(EZ_{(i)}) \leq \frac{i}{n}$ and $\Phi^{-1}(EZ_{(i)}) \approx \frac{i}{n+1}$. Also $EZ_{(n-i+1)} = -EZ_{(i)}$ and hence $\sum_{i=1}^{n} a_i = 0$. The test statistic is $r_n^2(a, x)$. Again, this choice (9) of a_i does not differ much from other choices, but p-values are available for this test. However, making the plot (4) with the choice (9) means comparing initially dependent observations against initially independent observations which does not sound very well justified.

There is no reason to assume that the p-values that are valid under the independence assumption would be valid also without the independence. (They could still be approximately valid, but we simply do not know.) So we try not to use them. Without p-values there is no standard argument to decide for example what would be the correct sample size, i.e. to decide, how close to 1 the value of r_n^2 should be for a given n. Note that for arbitrary increasingly ordered vectors the value of r^2 is *a priori* positive and already close to 1. P-values were useful also if the goodness of fit for samples of different sizes need to be compared.

To avoid the problems of which sample size to use and how to compare the goodness of fit for samples of different sizes we chose a rather pragmatic approach. We map r_n^2 against n and check whether it looks like that $r_n^2 \to 1$ or not, meaning that it is possible that the goodness of fit can improve when the sample size grows. Without the p-values we do not know whether the rate at which $r_n^2 \to 1$ is fast enough for the goodness of fit really to improve, but if there is clearly no convergence, we rule out the model.

The requirement of $r_n^2 \to 1$ is thus a necessary but not sufficient property. To obtain a sufficient criterion we compare the rate at which $r_n^2 \to 1$ of data to the rate calculated from simulated (uncorrelated) data, for which we know the null hypothesis hold. If the rate is the same, it means that this correlation test method cannot distinguish our traffic data from simulated data. For under the null hypothesis simulated uncorrelated data the clear linear shape in the corresponding Q-Q or N-Q plots is all the time present, and the fluctuations of the test statistic in the map $n \mapsto r_n^2$ are not easily seen from the corresponding Q-Q or N-Q plot. Simulations give us thus pairs of (n, r_n^2) that refer to good fit. However, since this is a test of global fit, it does not mean that the fit is good for example in the tails.

For a fixed sample size n the values of r_n^2 for two (or more) different data sets are comparable, hence this method gives also the opportunity to compare normal and log-normal models for the same data, or to compare different data for the same model.

V. Examples and results from data

Before presenting results from data we show a simulated example of the behavior of the map $n \mapsto r_n^2$ for Q-Q plots that compare two different samples. We simulated

four different uncorrelated traces, two from normal distributions with different parameters and two traces from uniform distributions, again with different parameters. The four thick grey lines of figure 9 show the comparison of both of the samples from normal distributions against both of the samples from uniform distributions. They seem to typically stay below the value 0.96, showing that without knowing the sample size n, even a large-looking value of r^2 does not tell anything about the goodness of fit. If the two compared samples are from distributions that are more similar than the normal and the uniform distributions, the maximum value of r_n^2 can be arbitrarily close to 1. But if the samples are from two statistically significantly different distributions (for example with positive Wasserstein distance), we do not expect that $r_n^2 \to 1$.

The two thin black curves of figure 9, coming very close to 1, compare the two simulated samples from normal distributions between themselves and the two simulated samples uniform distributions between themselves. They should give some evidence that if the two samples are from the same distribution, although with different parameters, we still expect to see that $r_n^2 \to 1$.

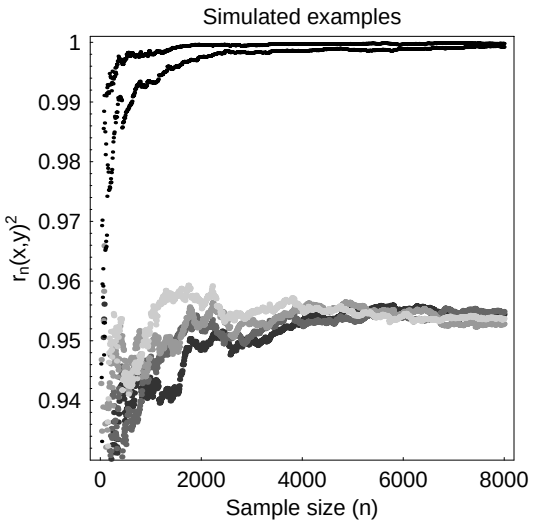

Fig. 9. Simulated examples of the behavior of the map $n \mapsto r_n^2$ for uncorrelated traces. The four thick grey curves compare samples from normal distributions against samples from uniform distributions. For this comparison the value of r_n^2 seems to stay below 0.96.

Next we will present examples from the real data. The thin light grey (most pale) lines in the following figures of this section that present the map $n \mapsto r_n^2$ are calculated from five different simulated uncorrelated data, for which the null hypothesis holds. For a fixed resolution Δ, the value $n \times \Delta$ gives directly the duration from the start times and it is also shown in the figures. The scale of the vertical axes is from 0.99 to 1, (except in the figures 18 and 17 in the subsection V-C). This scale makes the differences between traffic data and simulated data visible in all sample sizes that we use.

A. About the assumption of stationary time series

We make the two-sample test in the following way. We divided the 30 minute period into two 15 minute periods, and having the two start times we collected samples x_i starting from the beginning and y_i starting 15 minutes later. For each n we got vectors $x = (x_{(1)}, \ldots, x_{(n)})$ and $y = (y_{(1)}, \ldots, y_{(n)})$ and calculated the value $r_n(x, y)^2$ and plotted it against n. For a fixed Δ this can be done for up- and downstream data separately, and it is natural to compare the values r_n^2 for up- and downstream cases. Figures 10 and 11 show two examples of this procedure for two different resolutions. The lowest thick light grey curve in both figures represents the upstream values of $r_n(x, y)^2$ and the thick dark grey curve represents the corresponding downstream values.

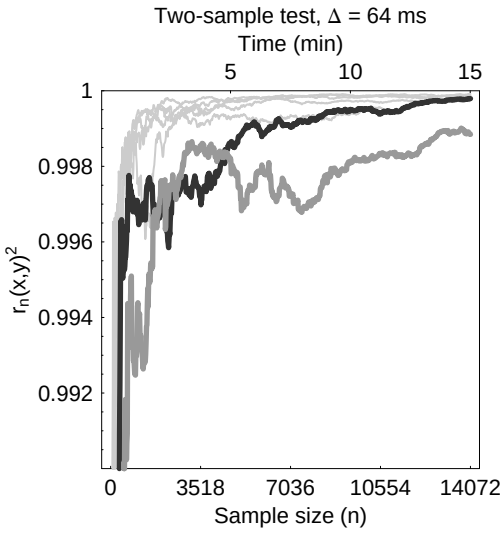

Fig. 10. The plot $r_n(x, y)^2$ against n. The darker (almost black) thick grey curve is calculated from the downstream data, thick light grey curve from upstream data.

The results for the last 30 minute period of our traffic data were that typically the downstream data gave better results than upstream data, but in both directions and in all resolutions from 64 ms to 4096 ms the value of $r_n(x, y)^2$ seems to converge towards 1, although slower than for the simulated uncorrelated data. The downstream rate was sometimes even comparable to simulated reference rate, whereas the upstream rate was almost always slower. In the downstream case we consider this result quite good.

The results of the same method for the first 90 minute period of the traffic trace, when divided into two 45 minute

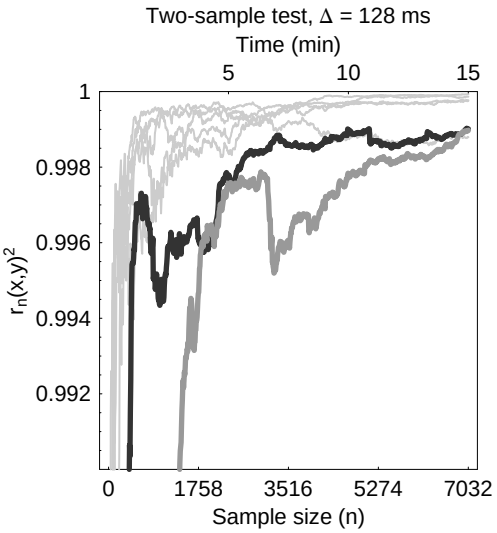

Fig. 11. Another example. For upstream data the value of $r_n(x,y)^2$ is typically all the observed time below the corresponding value of the downstream data.

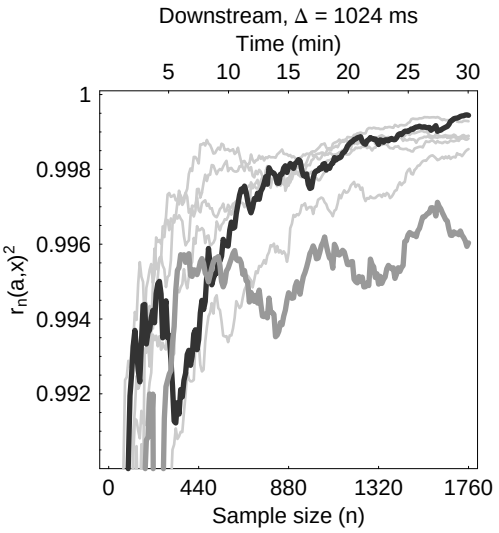

Fig. 12. For the original data it looks clearly that $r_n(a,x)^2 \to 1$ and the rate is all the time comparable to similated rate. See also figure 8.

periods, were similar but not so good, probably due to the fact that the level of vertical aggregation during the first 90 minute period of the traffic trace was essentially lower than after the increase at 17:00-17:10.

B. First time scale for good Gaussian approximation

In this subsection we search for the first time scale, i.e. level of horizontal aggregation, where the Gaussian approximation looks plausible for the data in the sense of our correlation test method.

Again we looked how the value of $r_n^2 = r_n(a,x)^2$ evolved in time, which now was the whole 30 minute period. We calculated simultaneously the value of r_n^2 for the original data x_i and for the log-transformed data $y_i = \log x_i$. Figures 12 and 13 show examples of these in the downstream direction, the lighter grey curve showing the log-transformed case.

We collected the results of the downstream case to table II below. For each resolution we show the better[3] model, log-normal or normal, and whether for this better model the value r_n^2 seems to tend towards 1 or not.

The conclusions for the downstream data are that the normal model is better than log-normal one, but only for the resolutions 1024 ms and higher we can assume that there is no better model, since only then the rate at which $r_n^2 \to 1$ was typically the same as simulated rates. For resolutions 128, 256 and 512 ms the linear shape in the corresponding N-Q plots was for some n pretty good, but for most n there were either clear deviations from the overall

[3]We do not claim that it is the best.

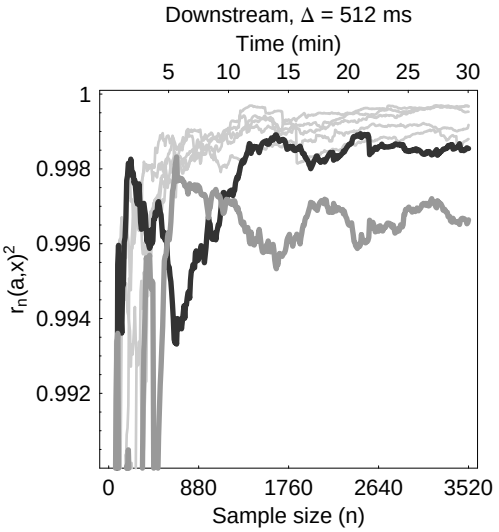

Fig. 13. In this case it is not so clear whether $r_n(a,x)^2 \to 1$, but the normal model is typically better.

TABLE II
DOWNSTREAM MODEL SELECTION.

Δ (ms)	Better model	$r_n^2 \to 1$?
64	normal	no
128	normal	possibly
256	normal	possibly
512	normal	possibly
1024	normal	yes
2048	normal	yes
4096	normal	yes

linear shape or fluctuations in the tails.

Table III below collects the same results in the upstream case. The situation here is different, as the log-normal model seems to fit better and also in these cases r_n^2 seems to tend towards 1. For resolution 512 ms and larger the rate was even comparable to simulated rates.

TABLE III
UPSTREAM MODEL SELECTION.

Δ (ms)	Better model	$r_n^2 \to 1$?
64	normal	no
128	normal	no
256	log-normal	yes
512	log-normal	yes
1024	log-normal	yes
2048	log-normal	yes
4096	log-normal	yes

Figures 14, 15 and 16 are examples of the upstream case.

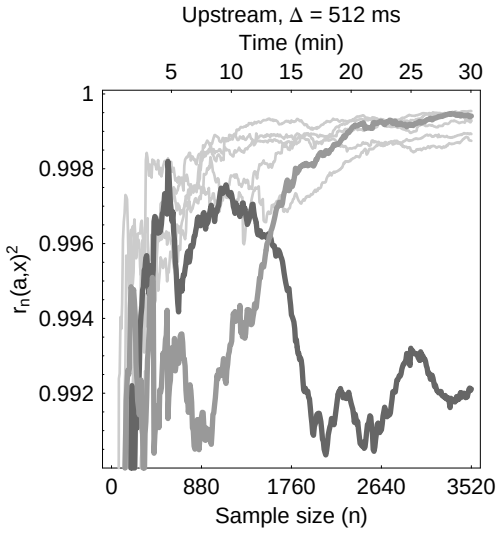

Fig. 14. The normal model seems to fit better for the first 10-15 minutes, but after that the log-normal model looks clearly superior.

A study of the empirical *skewness* per resolution also suggested that a positively skewed model (like log-normal distribution) fits essentially better for upstream traffic.

That upstream direction has this *subexponential* character can be explained as follows. Most of the upstream traffic consists of packets with no data. Especially, in the case of a single user, the contribution to the aggregate traffic consists mostly of relatively small bursts of few small packets and that happen occasionally.

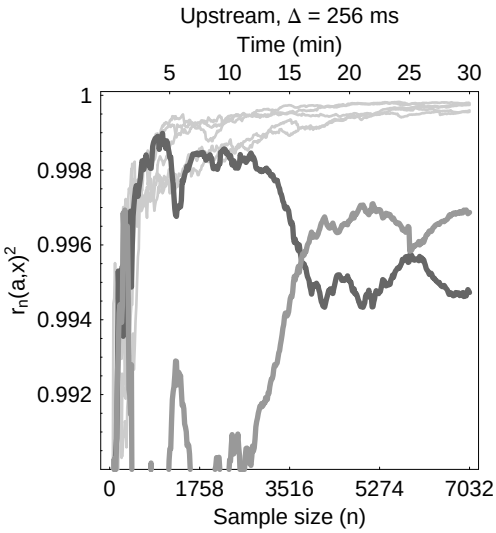

Fig. 15. Here again the normal model looks much better in the beginning, but the log-normal model seems finally the better.

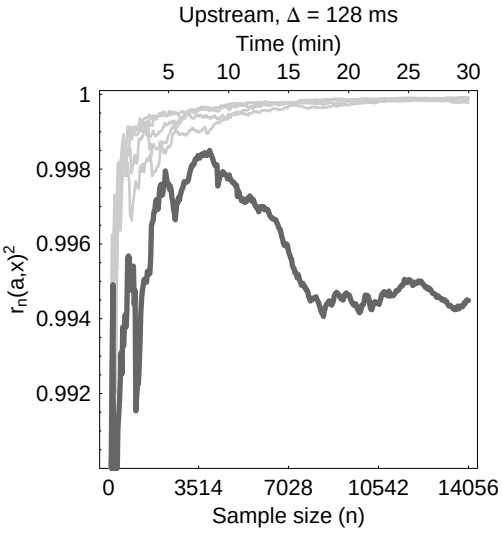

Fig. 16. In this resolution the value of r_n^2 for the log-transformed data stay all the time below 0.985, so it does not even show up in this figure. But the normal distribution is not a good model either!

When someone is sending data, for example mail with a large attachment, the burst consists of perhaps many large packets that are close, but not too close, to each others. The packets cannot be too close to each other since they come through an ISDN pipe with relatively slow rate when compared to the 100 Mb/s Ethernet rate. For a resolution large enough that a large burst of a single user fits completely to one time slot, the aggregate traffic rate obtains a remarkably high rate caused by this single contributor.

As the description of data, especially figure 2, shows, the number of those contributors that transmitted more

bytes than received is very small. When we observe the upstream traffic for long enough time, these contributors begin to show up, the upper tail becomes heavier and the marginal distribution, which in the beginning may look normal, becomes more asymmetric. This explains the phenomena of figures 14, 15 and 16 and shows that the level of aggregation in the upstream direction is not high enough to swallow an individual source that transmits data upstream.

There are also other subexponential distributions than the log-normal one, and there is no good explanation for the log-normal model. Moreover, the log-normal model is not very easy to analyze or use in practice. Recall that typically one arrives to the log-normal distribution when independent and strictly positive random variables are combined in a multiplicative way. A mixed model, a normal distribution for those who receive data and a separate model for those who transmit data would perhaps be correct, or good enough, in the upstream case.

C. *How much vertical aggregation would be needed in the smaller resolutions?*

It is natural to expect, because of CLT, that increasing vertical aggregation would decrease the Gaussian character threshold of each horizontal aggregation level. Although we use a single trace, we can study the effect of vertical aggregation in the following way. We assume that we are in the infinite capacity situation.

For a fixed resolution Δ we take k consecutive disjoint intervals of suitable length, overlap them, add up the bytes of each overlapping time slots in order to obtain a single sample, which is then tested.

Figures 18 and 17 show empirically that this method really does typically increase the value of r_n^2 for all n. In these figures we divided the 30 minute period to 15 periods of length 2 minutes, and made the artificially aggregated samples for the values $k = 2, 4, 8, 10, 12, 15$ and made the maps $n \mapsto r_n^2$.

To combine the amount of overlapping required for different resolutions we tested the overlapped data using the Shapiro-Francia normality test with risk level $\alpha = 0.01$. We used this test since we had to compare the goodness of fit of different size samples, and had to have exactly the same criteria in all of the cases. Anyhow, the whole idea is to get only rough results and this justifies the use of this test which, as was noticed before, relies rather heavily on the independence assumption. In this way we obtain a *threshold curve* of Figure 19 showing roughly the required mean traffic rate value for 8, 16, 32, 64 and 128 ms resolutions for the downstream direction required for a good Gaussian approximation in the sense of the Shapiro-Francia test. The figures 17 and 18 give evidence that the threshold curve is approximately correct.

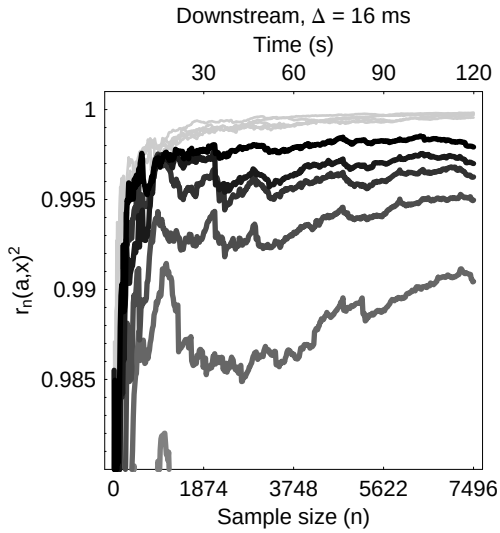

Fig. 17. The topmost (from the right edge) thick black curve represents the value $k = 15$, below it $k = 12$, then $k = 10$, $k = 8$ and $k = 4$.The curve with $k = 2$ show up in the lower left corner.

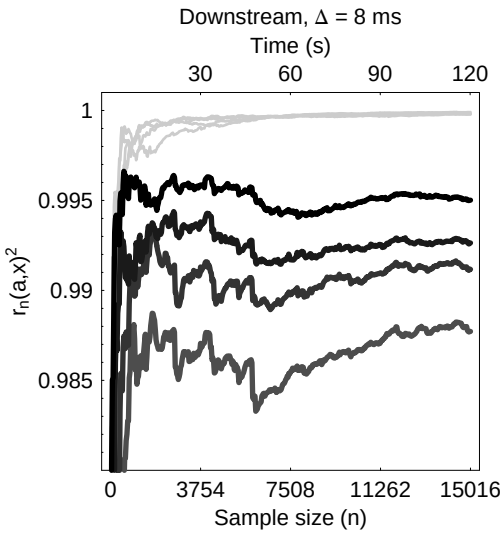

Fig. 18. For $\Delta = 8$ ms even the value $k = 15$ does not give nearby close to good fit. Note that the vertical scale is from 0.98 to 1.

Since we were limited in the test by sample size 5000, it was not meaningful to extend this method to smaller resolutions than 8 ms, for it covers only 40 seconds in time. The required mean traffic rate in the vertical axis of figure 19 is simply calculated by the transformation $k \mapsto k \times$ mean rate, where mean rate refers to the mean rate during the last 30 minute period of the trace. The edge of lighter grey area in figure 19 is the edge of the shaded area of the downstream picture of figure 6, transformed to the Mbps by assuming that each hypothetical

user contributes with μ, the empirical mean rate per user, calculated in the same way as in section III.

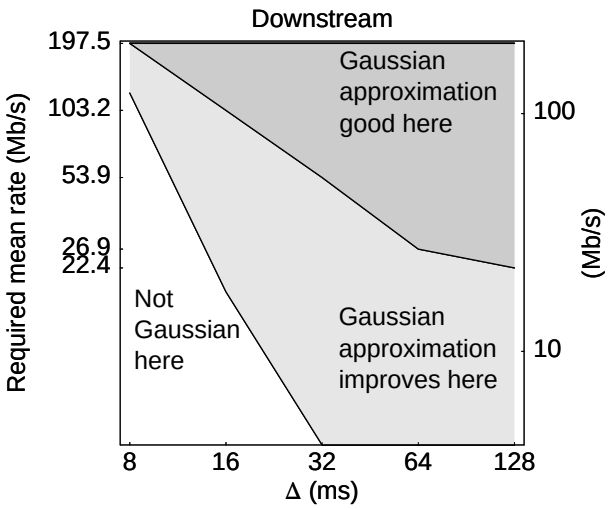

Fig. 19. Required mean traffic for being accepted Gaussian by the Shapiro-Francia test with significance level $\alpha = 0.01$.

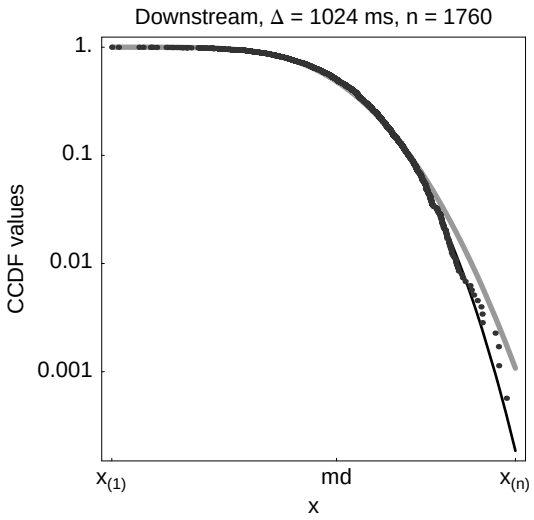

Fig. 20. The empirical complementary cumulative density function (CCDF) in log-log scale shows the typical situation that the extreme upper tail lies between normal and log-normal tail. Even in the case of very good fit like in this resolution. See also figures 8 and 12.

D. Distribution tails remain non-Gaussian

A problem of the Q-Q-plots and correlation tests in our context is that important differences in distribution tails are not well visible. In particular, the probability of observing a large queue is essentially determined by the probability that the amount of input traffic in a certain time interval Δ exceeds a large value.

Figure 20 shows that even when the global model fit is very good, the goodness of fit does not extend to the upper tail. For large enough samples the empirical upper tails were always heavier than normal tails but typically less heavy than log-normal tails. Moreover, the log-normal model was typically quite good in the extreme upper tail.

VI. Conclusions and further research topics

The results of our study can be summarised as follows:

1. We presented simple ideas to rule out cases, where the level of horizontal or vertical aggregation will not be sufficient for Gaussian approximation.
2. We presented also an elementary correlation test method, to look at the behavior of the map $n \mapsto r_n^2$, based on the linear correlation coefficient calculated from the corresponding Q-Q or N-Q plots, which does not assume independence of the observations.
3. The correlation test method can be used to rule out the normal model, and, in clear positive cases, to accept the approximative model. It gives some idea of how far or how close the normal approximation is. It can also be used in parallel with formal tests over Q-Q and N-Q plots, that are based on the independence assumption (see e.g. [13]), giving some idea how reliable the result of the particular test is.
4. The good Gaussian approximation in the sense of the correlation test method, when applied to aggregate TCP traffic data, may not extend to the upper tail of the distribution.
5. The tails of the example data are roughly log-normal, which is in line of the observations that the statistics of TCP traffic at small time scales resembles that of random multiplicative cascades.
6. For upstream direction of the example data, a positively skewed distribution (like the log-normal one) fits essentially better (globally) than the normal distribution.

The log-normality of the tails may seem fatal for using Gaussian models to estimate queue distributions. To state this in a provocative way: *only* the upper tails of the marginal distributions matter for the estimation of the small probabilities of large queues. But, on the other hand, the marginal distributions in question are those corresponding to the typical time of *build-up* of such a large queue — it is in fact that resolution whose Gaussian/non-Gaussian nature matters for queueing performance, and that resolution is, indeed, rather large.

Further work is needed, among other questions, to study the applicability of Gaussian queueing models at

different time resolutions (which in this case essentially means: different load/buffer combinations). In this context the desired hypothesis $H_0 : F \approx \Phi_{\mu,\sigma}$ might get some more precise content: the Gaussian approximation is good enough if the traffic "behaves like Gaussian" in some "meaningful queueing theory sense". For the moment though, this heuristic idea has no ground under its feet.

Can we make more use of the Wasserstein distance? For example, if we get a method to decide, that the probability that $F \in \mathcal{N}^\varepsilon$ is very high, where $\mathcal{N}^\varepsilon$ is the ε-neighborhood of the family of univariate normal distributions $\mathcal{N}$ in the Wasserstein metric, would that be useful in traffic theory?

The way we used Q-Q plots and correlation test method was to study a *global* fit of a model. However, Q-Q plots can also be used to study tails and a nice problem for further work is also whether the correlation test method also extends to the tails.

In backbone networks, the traffic rates can be very high, which is an argument for Gaussian modeling. However, the transmission speeds of invidual sources can also be very much higher than in our data. Moreover, as regards the tails, where rare high bursts dominate, even our low user speed results show that it is very hard to "swallow" those bursts by increasing vertical traffic aggregation.

Acknowledgement. We are grateful to Elisa Communications for allowing public research on data obtained from a commercial network, and to Kari Seppänen from VTT Information Technology for developing the actual measurement and data anonymization techniques used here.

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, Feb. 1994.

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

[3] I. Norros, "The management of large flows of connectionless traffic on the basis of self-similar modeling," in *1995 IEEE International Conference on Communications (ICC'95)*, Seattle, USA, 1995, pp. 451–455.

[4] N. G. Duffield and N. O'Connell, "Large deviations and overflow probabilities for the general single-server queue, with applications," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 118, no. 2, pp. 363–374, 1995.

[5] R.G. Addie, P. Mannersalo, and I. Norros, "Performance formulae for queues with Gaussian input," in *Teletraffic Engineering in a Competitive World. Proceedings of the International Teletraffic Congress — ITC-16*, P. Key and D. Smith, Eds., Edinburgh, UK, 1999, pp. 1169–1178, Elsevier.

[6] R. Addie, P. Mannersalo, and I. Norros, "Most Probable Paths and Performance Formulae for Buffers with Gaussian Input Traffic," *European Transactions on Telecomminications*, vol. 13, no. 3, pp. 183–196, 2002.

[7] I. Norros and J. Kilpi, "Gaussian traffic modelling for Differentiated Services," in *15th Nordic Teletraffic Seminar (NTS-15)*, Lund, Sweden, Aug. 2000, Lund University, pp. 219–230.

[8] P. Mannersalo and I. Norros, "Approximate formulae for Gaussian priority queues," 2001, ITC'17, Brasil.

[9] "COST-257: Impacts of new services on the architecture and performance of broadband networks," http://nero.informatik.uni-wuerzburg.de/cost/Final/.

[10] I. Norros and P. Pruthi, "On the applicability of Gaussian traffic models," in *The Thirteenth Nordic Teletraffic Seminar*, P.J. Emstad, B.E. Helvik, and A.H. Myskja, Eds., Trondheim, Aug. 1996, pp. 37–50, Norwegian University of Science and Technology.

[11] T. Konstantopoulos and S.-J. Lin, "Macroscopic models for long-range dependent network traffic," *Queueing Systems*, vol. 28, pp. 215–243, 1998.

[12] T. Mikosch, S. Resnick, H. Rootzén, and A. Stegeman, "Is network traffic approximated by stable Lévy motion or fractional Brownian motion?," *Annals of Applied Probability*, vol. 12, pp. 23–68, 2002.

[13] B. Brown and T. Hettmannsperger, "Normal scores, normal plots, and tests for normality," *Journal of the American Statistical Association*, vol. 91, no. 436, pp. 1668–1675, 1996.

[14] T. de Wet and J.H. Venter, "Asymptotic distributions of certain test criteria of normality," *South African Statistical Journal*, vol. 6, pp. 135–149, 1972.

[15] E. del Barrio, J.A. Cuesta-Albertos, and C. Matrán, "Contributions of empirical and quantile processes to the asymptotic theory of goodness-of-fit tests," *Test*, vol. 9, no. 1, pp. 1–96, 2000.

[16] J. Kilpi and I. Norros, "Call level traffic analysis of a large ISP," in *ITC Specialist Seminar on IP Traffic Measurement, Modeling and Management*, Monterey, CA, USA, Sept. 2000.

[17] E. del Barrio, Cuesta-Albertos J.A., C. Matrán, and J. Rodríquez-Rodríquez, "Tests of goodness of fit based on the L_2-Wasserstein distance," *Annals of Statistics*, vol. 27, pp. 1230–1239, 1999.

[18] J. Beran, *Statistics for Long-Memory Processes*, vol. 61 of *Monographs on Statistics and Applied Probability*, Chapman & Hall, 1994.

[19] S.S. Shapiro and M.B. Wilk, "An Analysis of Variance Test for Normality (Complete Samples)," *Biometrika*, vol. 52, no. 3/4, pp. 591–611, 1965.

[20] S.S. Shapiro and R.S. Francia, "An Approximate Analysis of Variance Test for Normality," *Journal of the American Statistical Association*, vol. 67, no. 337, pp. 215–216, 1972.