Ilmari Juva, Riikka Susitaival, Markus Peuhkuri, and Samuli Aalto. 2005. Traffic characterization for traffic engineering purposes: Analysis of Funet data. In Proceedings of the 1st EuroNGI Conference on Next Generation Internet Networks (NGI 2005), Rome, Italy, 18-20 April 2005, pages 404-411.

© 2005 IEEE

Reprinted with permission.

Traffic Characterization for Traffic Engineering Purposes: Analysis of Funet Data

Ilmari Juva, Riikka Susitaival, Markus Peuhkuri and Samuli Aalto
Helsinki University of Technology, Networking Laboratory
Email: {ajuva, rsusitai, puhuri, samuli}@netlab.hut.fi

Abstract— **For Internet traffic engineering purposes, it is important to characterize traffic volumes typically over 5-minute intervals. Based on measurements made in a local network at Lucent in winter 1999, Cao et al. [2] proposed a moving IID Gaussian model for the characterization of 5-minute traffic volumes, with a power-law relationship between the mean and the variance. In this paper we analyze novel measurements gathered from a 2.5 Gbps link in the Finnish university network (Funet) in summer 2004. We investigate the validity of the moving IID Gaussian model and the proposed mean-variance relationship when the measurement interval is varying from 1 second to 5 minutes. As a result, we find that the Gaussian assumption is much more justified with current core link rates. The mean-variance relationship seems, indeed, to follow a power-law with exponent approximately equal to 1.3 in our data set. However, the IID assumption concerning the standardized residual is not verified, but we find a clear positive correlation between adjacent 5-minute volumes, and only slightly weaker negative correlation for traffic volumes with distance 20-30 minutes. In addition, we demonstrate that the same phenomenon is already prevailing in the Lucent data set.**

I. INTRODUCTION

A basic concept in traffic engineering is a *traffic matrix* representing traffic demands between traffic sources (origins) and sinks (destinations). Thus, an entry in the traffic matrix tells the traffic volume (the number of bytes transferred) for a single OD pair over a specified time interval. In a standard IP network, these figures are not available but should be estimated, for example, from the link load measurements, which the Simple Network Management Protocol (SNMP) provides every 5 minutes.

The traffic volume for a single OD pair over a specified time interval is a random variable, in fact, a stochastic process varying randomly in time. Let X_n, $n = 1, 2, \ldots$, denote these traffic volumes in consecutive time intervals of length Δ. As mentioned above, for traffic engineering purposes we typically have $\Delta = 300$ s (i.e., 5 min). While there is a vast literature of papers considering traffic characteristics of modern data networks (starting from the well-known Bellcore measurements reports [5]) in small time scales, from microseconds to milliseconds and seconds, the longer time scales, from seconds to minutes and hours, have not been explored that much.

Vardi [9] proposes a Poisson model, where X_n are Poisson distributed. In particular, this would imply that the variance equals the mean,

$$D^2[X_n] = E[X_n].$$

As Cao et al. [2] notes, this is not scale-invariant. Using, for example, bits per second (instead of bytes per second) leads to another variable $Y_n = 8X_n$, for which

$$D^2[Y_n] = 64D^2[X_n] \neq 8E[X_n] = E[Y_n].$$

Cao et al. [2] propose a *moving IID Gaussian model* with the following relationship between the variance and the mean,

$$D^2[X_n] = \phi E[X_n]^c. \tag{1}$$

In this model the exponent c is scale-invariant while the factor ϕ depends on the unit used. This can bee seen from our previous example, where $Y_n = 8X_n$. In this case,

$$D^2[Y_n] = 64D^2[X_n] = 64\,\phi E[X_n]^c = \phi' E[Y_n]^c,$$

where $\phi' = 8^{2-c}\phi$. Based on measured link counts from a local network at Lucent with time scale $\Delta = 300$ s, Cao et al. conclude that "a quadratic power law ($c = 2$) is more reasonable than a linear law ($c = 1$)." The moving IID Gaussian model captures the time-varying nature of real traffic as demonstrated by typical diurnal patterns. In this model the standardized residuals,

$$\frac{X_n - E[X_n]}{D[X_n]},$$

are assumed to be independent and identically distributed according to a standard normal marginal distribution. Cao et al. describe the agreement of the measured data with the Gaussian assumption as sufficient. The IID property is verified only visually by plotting the sample-standardized residuals as a time series.

Medina et al. [6] report that while the power-law relationship (1) seems to hold, the exponent c varies remarkably from one link to another within bounds $c \in [0.5, 4.0]$. These observations are related to data collected from a tier-1 backbone (i.e., from the Internet core) with time scale $\Delta = 1$ s. Note that the time scale is very different from that used in [2].

Morris and Lin [7] find a linear relationship between the variance and the mean ($c = 1$) for Web traffic. They base the statement on traffic traces from Harvard's campus network (100 Mbps Ethernet) and a local network at Lucent (Ethernet). The time scale used in this study is even shorter, $\Delta = 0.1$ s.

In a recent study, Gunnar et al. [3] confirm the validity of the mean-variance relationship (1) and give values $c = 1.5$ and 1.6 based on data traces from a global operator's backbone with $\Delta = 300$ s.

In this paper we investigate the validity of the moving IID Gaussian model with the mean-variance relationship (1) when the time scale Δ is varying from 1 s to 300 s. Our measurement data comes from Finnish university network Funet. More specifically, our measurement data consists of link counts with 1-second resolution taken from an STM-16 link (2.5 Gbps) over a 13-day period in summer 2004. In addition, we use the same data set as in [2] to demonstrate the similarities/differences in the characteristics of traffic coming from two very different environments. Throughout the paper we use bits per second (bps) as the unit of the link count. Thus, for example, if the measurement gives the byte count of 300 Mbytes over a 5-minute period, the corresponding link bit count will be $X_n = 8$ Mbps.

The rest of the paper is organized as follows: Section II studies analytically why the linear mean-variance relationship may be reasonable. Section III describes the data sets used. Section IV tests whether the Gaussian assumption is valid for Funet data. The purpose of Section V is to investigate the IID assumption of standardized residuals by autocorrelation function. Section VI analyzes the mean-variance relation of our data set. Finally, Section VII concludes the paper.

II. PRELIMINARY

In this section we give some simple and intuitive arguments why the mean-variance relationship (1) with $c = 1$ might be valid in current backbones.

Packet-level Bernoulli model: First we make a simple investigation at packet level. Consider a link of capacity b (e.g., $b = 2.5$ Gbps) loaded by traffic with average rate m (e.g., $m = 0.5$ Gbps). In the shortest possible time scale the link is either fully utilized, during packet transmissions, or idle. Thus, R_p, which refers to the traffic rate at this packet time scale, is a scaled Bernoulli random variable with distribution

$$P\{R_p = b\} = \frac{m}{b}, \qquad P\{R_p = 0\} = 1 - \frac{m}{b},$$

and satisfying the following mean-variance relationship

$$D^2[R_p] = m(b - m) = bE[R_p] - E[R_p]^2. \tag{2}$$

For lightly loaded link ($m \ll b$), we have the following approximative mean-variance relationship:

$$D^2[R_p] \approx bE[R_p]. \tag{3}$$

To get some idea of the time scale, we note that it takes about 5 microseconds (μs) to transmit a typical IP packet of length 1500 bytes with link rate $b = 2.5$ Gbps.

Flow-level Poisson model: Enlarging the time scale to the flow level, consider an overdimensioned backbone link of Gbps-size. Single flows coming from Mbps-size access networks do not find the backbone link as a bottleneck. According to recent studies [1], flow arrivals in such a link constitute a Poisson process, which in a longer time scale is surely time-varying but in a shorter time scale can be considered as time-homogeneous. On the other hand, it is widely accepted that the flow sizes follow a heavy-tailed distribution, such as

Pareto with shape parameter $\alpha \in (1, 2)$. These together imply an $M/G/\infty$ model according to which the instantaneous number of active flows, N_f, follows a Poisson distribution with property

$$D^2[N_f] = E[N_f].$$

If all the flows had the same bottleneck bandwidth h, the total instantaneous traffic rate at flow level would be $R_f = N_f h$ satisfying the following mean-variance relationship

$$D^2[R_f] = h^2 D^2[N_f] = h^2 E[N_f] = hE[R_f]. \tag{4}$$

As an illustrating example, we mention that it takes 8 seconds to transfer a file of size 1 Mbyte with an access rate $h = 1$ Mbps.

The model can easily be generalized by allowing each flow i an IID bottleneck bandwidth H_i. In this case, the mean-variance relationship reads as

$$D^2[R_f] = E[R_f]E[H](1 + C^2[H]), \tag{5}$$

where $C^2[H]$ refers to the squared coefficient of variation,

$$C^2[H] = \frac{D^2[H]}{E[H]^2}.$$

III. DATA SETS
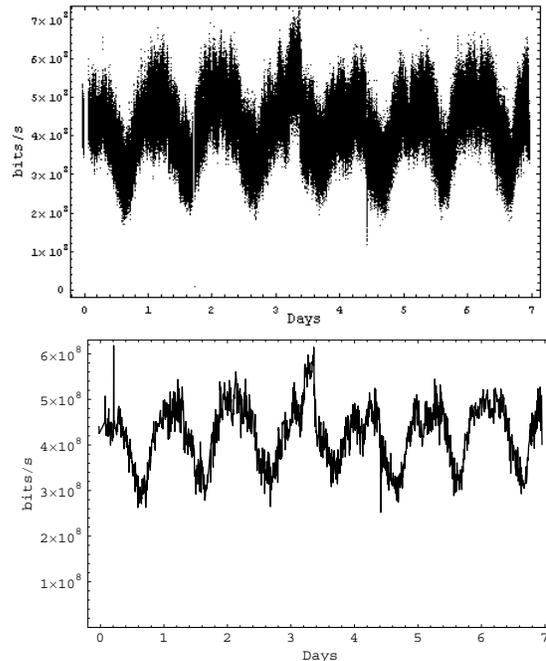
Fig. 1. Trace of the Funet data set (top, $\Delta = 1$ s) and 5 minutes aggregated traffic (bottom, $\Delta = 300$ s), over one week.

In this section we describe our measurement data sets, one from Finnish university network, Funet, and another from a local network at Lucent [2]. For the new Finnish traffic traces, we also describe the methodology needed to create these data sets from the raw traffic traces.
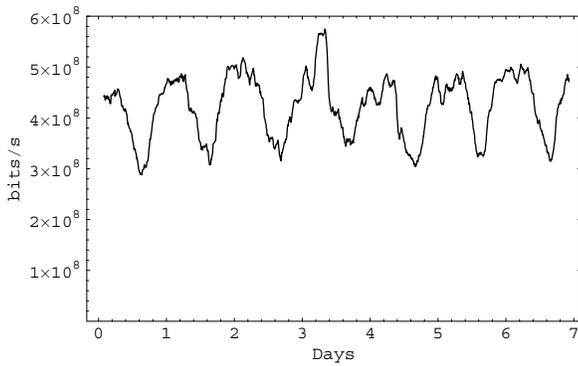
Fig. 2. The moving sample-average m_n^Δ of Funet trace, $\Delta = 300$ s.
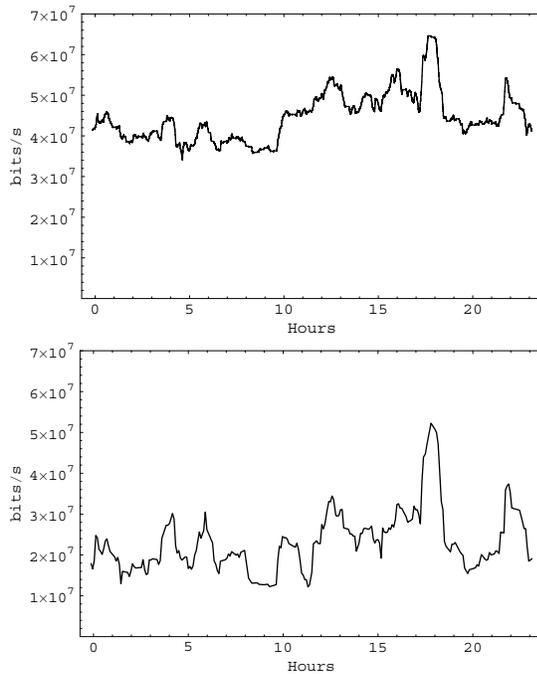
Fig. 3. The moving sample-standard-deviation s_n^Δ of Funet trace with $\Delta = 1$ s (top) and $\Delta = 300$ s (bottom).

Fig. 4. The sample-standardized residual z_n^Δ of Funet trace , with $\Delta = 1$ s (top) and $\Delta = 300$ s (bottom).

Fig. 5. The cumulative number of bits counted over 5-minute interval.

A. Funet measurement methodology

Funet traces were captured between csc0-rtr and helsinki0-rtr[1] from 2.5 Gbps STM-16 link using Endance DAG 4.23 cards. The IP addresses on captured packet headers were anonymized preserving prefix, and the headers were stored to disk using flow-based compression [8]. Captured traffic was transferred once an hour to an analysis machine where statistics were calculated. Part of the traces were archived for later analysis, but not all because of large volume of data (about 10 Mbps average). For this study, bytes transferred each second were calculated.

In our measurement data set, TCP accounts more than 98 % of bytes transferred. During daytime 10-20 % of TCP traffic is

[1]For details about Finnish university network (Funet), see http://www.csc.fi/suomi/funet/verkko.html.en

HTTP. There exists also considerable amount of peer-to-peer traffic. Part of data points were missing because of transient errors in data analysis.

B. Funet data

Original data: The Funet data set consists of link counts (in bits) measured in one second intervals over two periods; first one from 3am June 29th to 2pm July 6th, and the second from 10am August 3rd to 12am August 10th of 2004 local time. Denote this original measurement data by $x = (x_t; t = 1, 2, \ldots, T)$, where x_t refers to the measured link bit count at time t seconds (from the beginning of the measurement period). These link bit counts for the first 7 days are shown as a function of time at the top of Figure 1. From this figure, we can see that the traffic has a strong diurnal pattern. The traffic rates during the day time (around 500 Mbps) are larger than during the night time (around 300 Mbps). Different days look
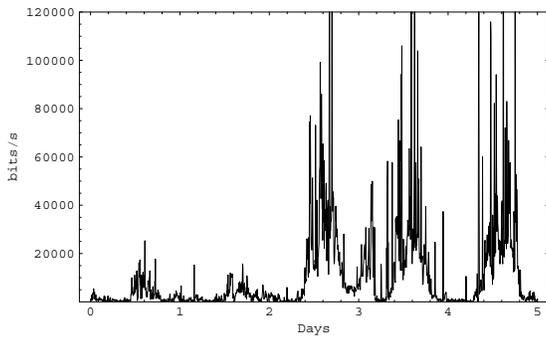
Fig. 6.   Typical OD-pair trace for Lucent data

very much the same. Only one day differs clearly from the others having a slightly higher peak rate at the busy period.

Link bit counts over periods of length Δ*:* For each time scale $\Delta = 1,...,300$ s investigated, we created the corresponding time series of link bit counts $x^\Delta = (x_n^\Delta; n = 1, 2, \ldots, T/\Delta)$ by defining

$$x_n^\Delta = \frac{1}{\Delta} \sum_{t=n\Delta+1}^{(n+1)\Delta} x_t.$$

Trace for a SNMP-like traffic, for which $\Delta = 300$ s, is shown at the bottom of Figure 1.

Traffic components: As in [2], we separate different components from the link bit counts x_n^Δ,

$$x_n^\Delta = m_n^\Delta + s_n^\Delta z_n^\Delta,$$

where m_n^Δ refers to the moving sample-average, s_n^Δ to the moving sample-standard-deviation, and z_n^Δ to the sample-standardized residual. Based on Figure 1, the averaging period was chosen to be (about) 1 hour. Thus,

$$m_n^\Delta = \frac{1}{3600/\Delta + 1} \sum_{k=n-1800/\Delta}^{n+1800/\Delta} x_k^\Delta$$

and

$$s_n^\Delta = \sqrt{\frac{1}{3600/\Delta + 1} \sum_{k=n-1800/\Delta}^{n+1800/\Delta} (x_k^\Delta - m_k^\Delta)^2}$$

The moving sample-average, m_n^Δ, is depicted as a function of time in Figure 2 for the first 7 days. The moving average for a given time moment does not essentially depend on the aggregation and thus only the curve for $\Delta = 300$ s is depicted. The moving sample-standard-deviation is depicted for one day period at two time-scales, $\Delta = 1$ s and $\Delta = 300$ s, in Figure 3. The standard deviation seems to be time-dependent. In addition, it decreases as the aggregation level increases. Finally, the remaining component, the standardized random fluctuation z_n^Δ is shown for the same time period in Figure 4. At the short time scale ($\Delta = 1$ s) standardized residuals seem to be random noise, which is not a case for 5 minute aggregates ($\Delta = 300$ s).
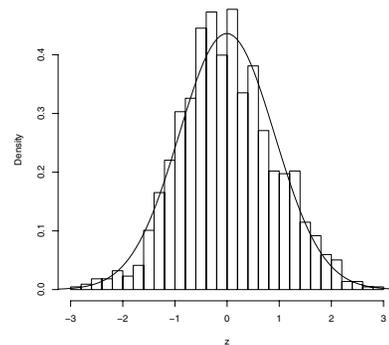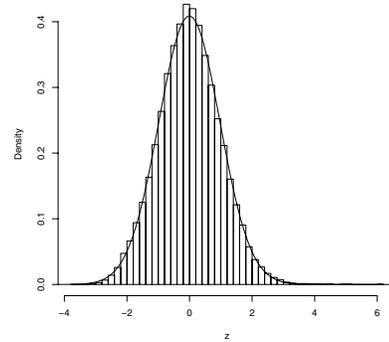
Fig. 7.   Histograms of Funet data with $\Delta = 1$ s (top) and $\Delta = 300$ s (bottom) against the normal distribution density function.

Variation within 5-minute intervals: The cumulative bit count over a 5-minute is shown in Figure 5. The cumulative bit count within this 5-minute period seems to be surprisingly linear. In [4] similar counts are presented for both 15 years and 2 years old traces and they indicate that traffic has been much more non-stationary in this time scale.

C. Lucent data

The Lucent data set contains data from two routers, which have 12 and 5 nodes, respectively, connected to them with one-way links. The measurements consist of link bit counts over 5 minute periods implying that $\Delta = 300$ s for this data. In Figure 6 a typical example of the OD-pair traces is shown. The trace is over five days, meaning 1440 values of five minute averages. In some of the OD-pairs the traffic volumes are much smaller over the first two days indicating different traffic rates in the weekends as compared to working days. Another obvious difference to the Funet data is variability, traffic rate is many times close to zero in the Lucent routers.

IV. TESTING GAUSSIAN ASSUMPTION

In this section we test whether a Gaussian assumption is valid for the data sets used. We concentrate on the stochastic component, the standardized residual z_n^Δ as defined in Section III, and study measurements of one second interval as well as the five minute aggregates.
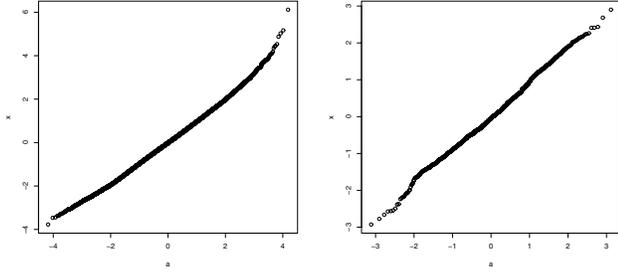
Fig. 8. N-Q tests comparing the Funet-data with $\Delta = 1$ s (left) and $\Delta = 300$ s (right) against the normal distribution.

Fig. 9. Autocorrelations of Funet data in one second measurement interval ($\Delta = 1$ s).

In Figure 7 the histograms of the Funet data are shown comparing them against density function of the normal distribution. For the one second time scale the Gaussian density function follows the data nicely. For the five minute aggregates the curve does not follow the histogram as closely, but there is a reasonably good fit.

A good way to evaluate the appropriateness of the Gaussian assumption is the normal quantile (N-Q) plot. The original sample vector x is ordered from the smallest to the largest and plotted against vector a, which is defined as

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

where Φ is the cumulative distribution function of the normal distribution. The values given for a are between 0 and 1, so that the vector a contains the normal quantiles, having values from approximately -3 to 3. If the considered data follows the normal distribution, the plot should be linear. Goodness of fit with respect to this can be calculated by the linear correlation coefficient r, and the value r^2 is used as a measure of the fit.

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

The N-Q plots shown in Figure 8 confirm the strong Gaussianity observed in the histograms. The r^2-values are 0.999 and 0.996 for the one second measurements and the five minutes aggregates respectively. Thus we can conclude that the normal approximation cannot be rejected based on this.

In [2] the N-Q plot for the Lucent data set is given. The distribution of that data set has heavier tails than the normal distribution and also high peaks around the mean, which causes visible concavity for the N-Q plot. This seems to be the case for most OD pairs in the data set. Cao et al. conclude that the fit is still sufficient for the normal-approximation to be used.

V. AUTOCORRELATION

In this section we study the autocorrelations of the data with respect to different aggregation intervals. For the observations truly to be considered IID, there should not be any significant autocorrelations in the stochastic component z_n^{Δ}.
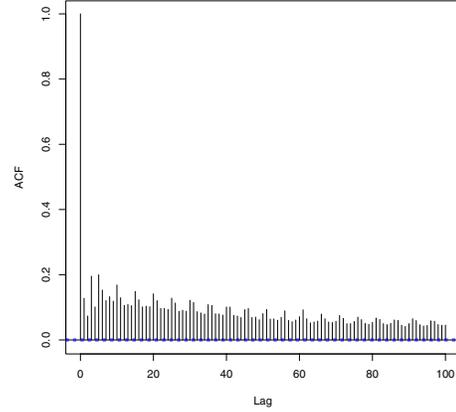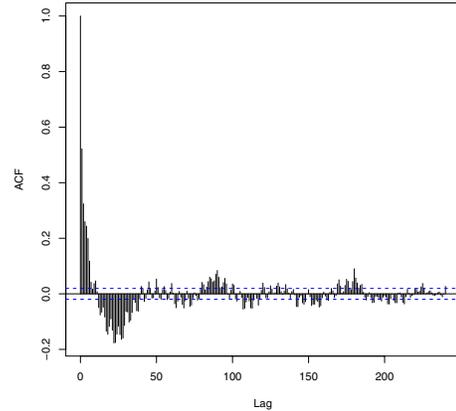
Fig. 10. Autocorrelations of Funet data from one minute aggregate measurements ($\Delta = 60$ s).

Figure 9 shows autocorrelation in the one second measurement intervals for one week of the Funet data. Clearly there are positive autocorrelations, meaning dependency between consecutive measurements. In the case of one minute aggregates for the Funet data, depicted in Figure 10, we notice significant positive values up to a lag of little over five minutes, and then a set of negative autocorrelation values after that is clearly observable. In Figure 11 the autocorrelation of the five minute aggregates of the Funet data as well as the autocorrelation of a typical Lucent data OD-pair are shown. The autocorrelation of the Funet data obviously corresponds nicely to the behavior of the one minute aggregates of the same data set. Comparing Lucent and Funet five minute measurements, a noticeable result is that the autocorrelation function of these two very different data sets seem to be surprisingly similar. In both cases it is not until a lag of more than thirty minutes that there is not any significant autocorrelation.
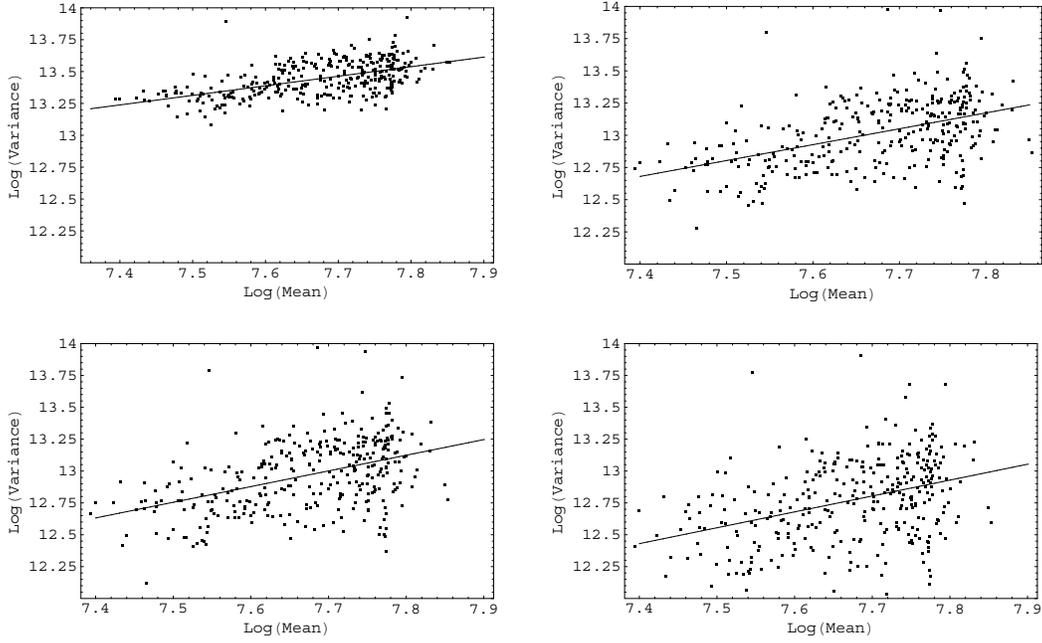
Fig. 12. Mean Variance relation scatter plot and best functional fit for Funet data for sample intervals of one second (top left), 30 seconds (top-right), 60 seconds (bottom left) and 300 second (bottom right).

VI. MEAN-VARIANCE RELATION

In this section we consider whether the mean-variance relationship defined in (1) applies to our data sets. It is important to separate between relations over a set of OD-pairs as studied in [2] and [3] and on the other hand relation within an OD-pair or link over time as studied in [6]. We concentrate mainly on the latter.

By taking the logarithm of (1), we see that

$$\log D^2[X_n] = c \log E[X_n] + \log \phi.$$

Thus, in the logarithmic scale, the exponent c is a linear coefficient, and its value may be obtained from the slope of the line that fits the scatter best in the least square sense. Indeed there is a noticeable linear dependence in the figures to be shown below.

Funet data: The mean-variance relation is studied by dividing the data into non-overlapping one hour periods and calculating the sample mean and sample variance for each of the periods. The variances as a function of the mean at the log-log scale, along with the lines depicting the best functional fit, are shown in Figure 12 for original one second sample interval, and for aggregated measurements with $\Delta = 30, 60, 300$ s. The estimates of c and ϕ for different aggregation levels are given in Table I. When the aggregation level is short, exponent c is less than one. Increasing the level of aggregation also increases c until it reaches $c = 1.25$ for the five minute aggregates. Correspondingly, the estimate for ϕ decreases exponentially.

Lucent data: In the Lucent data we have the opportunity to study mean-variance relation both within an OD-pair or link but also between all the OD-pairs. In [2] a subset of the Lucent

TABLE I
THE ESTIMATES FOR c AND ϕ WITH DIFFERENT LEVEL OF AGGREGATION.

Δ	c	ϕ
1	0.75	8.8
2	0.92	7.2
3	1.01	6.42
4	1.07	5.82
5	1.11	5.44
10	1.19	4.71
30	1.23	4.26
60	1.23	4.23
300	1.25	3.89

data set is studied for the latter kind of relation. They find that setting the power $c = 2$ is a good fit for the data set and our studies yielded similar result for the whole Lucent set.

Here, we consider the relation within an OD-pair. As in the previous section, each OD-pair is broken down into one hour periods and sample mean and sample variance are calculated for each of these periods. In Figure 13 the relation is shown along with the lines depicting cases $c = 1$, $c = 2$ and the best functional fit, which in this case is $c = 1.73$. However, the value of c differs from OD-pair to OD-pair in the rather wide range of $[1.2, 4.1]$ for the 20 largest OD-pairs considered here. A result similar to that was found by Medina et al. [6].

In Figure 14 sample points of these one hour periods are plotted for all OD-pairs in the same figure. For this the best fit is achieved with $c = 1.96$. This, of course echoes as much the relation over the OD-pairs as the relation within the OD-pairs.

Fig. 13. Mean-variance relation of one hour periods over a single OD-pair form the Lucent data set
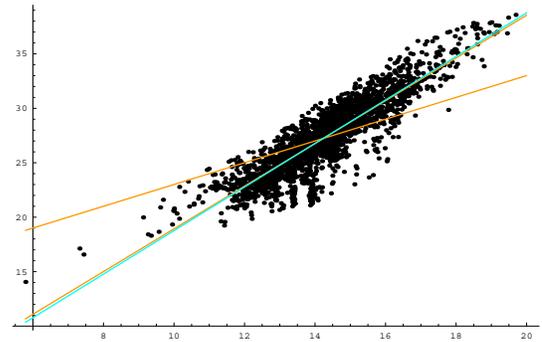
Fig. 14. Mean-variance relation of the Lucent data set with one hour periods over all the OD-pairs combined

Fig. 11. Autocorrelation of Funet data with five minute aggregation measurements (top, $\Delta = 300$ s) and autocorrelation of a typical OD-pair from the Lucent data on the right (bottom).

VII. CONCLUSION

In this paper we studied the validity of two common traffic engineering assumptions concerning the OD counts: the Gaussian IID model of observed measurements, and the functional relation between mean and variance.

In section IV the normal distribution was found to be a next to perfect fit for the distribution of measurements on one second interval. For the SNMP-like five minute aggregates the fit is not quite as good but still very satisfactory. We can conclude that the Gaussian model is justifiable. In section V the autocorrelation functions of the traces were studied. In both Lucent and Funet data we found significant values of autocorrelation for lags as long as half an hour. Strict IID assumption is thus not valid, even for five minute aggregated measurements, let alone for shorter intervals. It is left for further research to find a better model reflecting these observed correlations.

For the Lucent data we found out that regarding the mean-variance relation within an OD-pair, the power law parameter c varies between OD-pairs. This result is in line with those obtained by Medina et al. [6] stating that the parameter varies from link to link. For the Funet data, we learned the value of the parameter to be $c = 1.25$ for the five minute measurement intervals, but smaller for shorter intervals.

REFERENCES

[1] C. Barakat, P. Thiran, G. Iannaccone, C. Diot, and P. Owezarski, Modeling Internet backbone traffic at the flow level, IEEE Transactions on Signal Processing, Vol. 51, No. 8, pp. 2111–2124, 2003.
[2] J. Cao, D. Davis, S. V. Wiel, and B. Yu, Time-varying network tomography, Journal of the American Statistical Association, Vol. 95, pp. 1063–1075, 2000.
[3] A. Gunnar, M. Johansson, and T. Telkamp, Traffic matrix estimation on a large IP backbone – A comparison on real data, IMC'04, Taormina, Italy, 2004.
[4] T. Karagiannis, M. Molle, M. Faloutsos, and A. Broido, A nonstationary Poisson view of Internet traffic, in IEEE Infocom, Hong Kong, 2004.
[5] W. Leland, M. Taqqu, W. Willinger, and D. Wilson, On the Self-Similar Nature of Ethernet Traffic (Extended Version), IEEE/ACM Transactions on Networking, Vol. 2, No. 1, pp. 1–15, February 1994.
[6] A. Medina, N. Taft, K. Salamatian, S. Bhattacharyya, and C. Diot, Traffic matrix estimation: Existing techniques and new solutions, in ACM SIGCOMM, Pittsburg, USA, 2002.

[7] R. Morris and D. Lin, Variance of aggregated web traffic, in IEEE Infocom, Tel Aviv, Israel, 2000.

[8] M. Peuhkuri, A method to compress and anonymize packet traces, in IMW'01, San Francisco, USA, 2001.

[9] Y. Vardi, Network tomography: Estimating source-destination traffic intensities from link data, Journal of the American Statistical Association, Vol. 91, No. 433, pp. 365–377, 1996.