# Characteristics of Origin-Destination Pair Traffic in Funet

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

## Abstract

*In this paper we analyze measurements gathered from a 2.5 Gbps link in the Finnish university network (Funet) in 2004. The traffic is broken down into origin-destination (OD) pair components based on source and destination IP address. We study the traffic characteristics of these components, and identify four typical representative OD pairs. For these pairs we investigate the validity of a moving IID Gaussian model. We find that the statistical properties of these OD pairs differ significantly from each other, with only some of them close to Gaussian. The OD pairs are also found to have some cross-correlation between each other, contradicting an often made assumption about OD pair independence. Furthermore, the existence of a mean-variance relation between the OD pairs is studied. We find that there is a relation between mean and variance, but for some periods of time it is rather weak.*

## 1 Introduction

In IP networks the Simple Network Management Protocol (SNMP) provides link load measurements periodically, with a typical period of $\Delta = 5$ minutes. The main features of such traffic samples are i) aggregation in time (a measurement period of a couple of minutes), ii) aggregation in space (link traffic originating from a multitude of active flows), iii) a network wide view (all links measured).

Let $y_j$, with $j = 1, \ldots, J$, denote the results of such link load measurements, i.e., $y_j$ tells the number of bits transferred over link $j$ during a certain measurement period. Traffic $y_j$ consists of the traffic of those origin-destination (OD) pairs that use link $j$. The primary (but typically unknown) traffic demands $x_k$, with $k = 1, \ldots, K$, describe the traffic volumes (bit counts) over the measurement period for each OD pair $k$. Assuming that the routes do not change during the measurement period, there is a linear relationship between the vector of measured link counts, $\boldsymbol{y} = (y_j; j = 1, \ldots, J)$, and the vector of unknown traffic demands, $\boldsymbol{x} = (x_k; k = 1, \ldots, K)$:

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x},$$

where $\boldsymbol{A}$ refers to the $J \times K$ routing matrix with elements $A_{jk} = 1$ (0) if the traffic of OD pair $k$ is (not) routed via link $j$. As well known, inferring the unknown traffic demands $\boldsymbol{x}$ from the measured link loads $\boldsymbol{y}$ is a highly under-constrained problem. Many different methods have been developed to solve this ill-posed inverse problem. See the recent reviews and evaluations presented in [1, 2, 3, 4].

The traffic demands as well as the link counts obviously vary in time. Let $y_{nj}$ denote the link count related to link $j$ and measured over measurement period $n$, with $n = 1, \ldots, N$. The corresponding traffic demands are denoted by $x_{nk}$. Furthermore, let $\boldsymbol{Y}$ denote the corresponding $N \times J$ matrix and $\boldsymbol{X}$ the corresponding $N \times K$ matrix.

Lakhina et al. [5] present an interesting structural analysis of the traffic demand matrix $\boldsymbol{X}$. By applying Principal Component Analysis (PCA), they demonstrate that each OD flow $(x_{nk}; n = 1, \ldots, N)$ can be well approximated by a linear combination of a small number of so called eigenflows. In addition they observe that these eigenflows fall into three categories: *deterministic* eigenflows exhibiting strong diurnal periodicity, *spike* eigenflows with clear outliers, and *noise* eigenflows with a nearly Gaussian marginal distribution. These observations are based on sampled flow data from two different backbone networks, one from Europe and one from the US.

Soule et al. [6] observe from sampled flow data collected from a commercial Tier-1 backbone that large and medium size OD flows contain (at least) two sources of variability. This coincides with the eigenflows found by Lakhina, as the OD pairs have a deterministic cyclostationary diurnal patterns along with noisy fluctuations with zero mean.

These and other novel analyzes demonstrate that the large OD flows within backbone networks tend to have clear diurnal patterns, while small OD flows lack this property. This is in line with older measurements of smaller flows within local area networks, see for example [7].

Another approach is to consider the OD flows $(x_{nk}; n = 1, \ldots, N)$ as realizations of corresponding stochastic processes $(X_{nk}; n = 1, \ldots, N)$. Cao et al. [7] proposed a *moving IID Gaussian model*, consisting of a deterministic term $E[X_{nk}]$ capturing the possible cyclo-stationary diurnal pattern and a randomly fluctuating term $D[X_{nk}]Z_{nk}$,

$$X_{nk} = E[X_{nk}] + D[X_{nk}]Z_{nk}, \tag{1}$$

where the standardized residuals,

$$Z_{nk} = \frac{X_{nk} - E[X_{nk}]}{D[X_{nk}]}, \tag{2}$$

are assumed to be independent and identically distributed according to a standard normal distribution with zero mean and unit variance. Note that there is neither temporal nor spatial correlation among residuals $Z_{nk}$. In addition, Cao et al. included in their model a specific *mean-variance relationship*,

$$D^2[X_{nk}] = \phi E[X_{nk}]^c. \tag{3}$$

The exponent $c$ is scale-invariant while the factor $\phi$, naturally, depends on the data unit used.

The validity of this moving IID Gaussian model has been tested against different data sets. Cao et al. themselves use measured link counts $y_{nj}$ from a local network at Lucent with measurement period $\Delta = 300$ s. They describe the agreement of the measured data with the Gaussian assumption as "sufficient", but from an outsider's point of view the evidence they present is not that convincing. In addition, Cao et al. verify the IID property only visually by plotting the sample-standardized residuals as a time series.

Regarding the mean-variance relationship (3) over all OD pairs, conclusions of its validity vary. Cao et al. conclude that a quadratic power law is a reasonable fit. Gunnar et al. [2] as well as Medina et al. [1] also confirm the validity of the relation, while Soule et al. [6] conclude that it cannot be confirmed based on their study. Concerning the relation over time in one OD pair or link, results in literature are more in line with each other. It is found in [1, 6, 8] that the parameter values as well as the validity of the relation vary dramatically from OD pair to OD pair.

In [8] we investigate the validity of the moving IID Gaussian model against Finnish university network (Funet) link count measurements, varying the measurement period from 1 second to 5 minutes. We find that the Gaussian assumption is justified for the aggregated link traffic. However, the IID assumption seems to be questionable as there is a clear positive correlation between adjacent 5-minute aggregates. Regarding the mean-variance relationship, our estimate for the power-law constant $c$ increases as a function of the measurement period $\Delta$, with the value $c = 1.25$ corresponding to measurement period $\Delta = 300$ s.

In the present paper we continue the analysis of the Funet data. This time we split a 24-hour packet trace, captured from a single link, into separate OD flows. The origin of a packet is determined from the source IP address with a 10-bit mask, i.e. the first 22 bits of the address are used for the classification. The destination is determined correspondingly from the destination IP address of the packet. First we make a rough classification of some major OD flows identifying four different OD flow types. The representatives of these four OD flow types are then chosen to a more detailed analysis. In this analysis we investigate the validity of the moving IID Gaussian model. In addition, we try to find out possible cross-correlations, and also study the mean-variance relation in this data set.

The rest of the paper is organized as follows. In the next section we will review the measurement methodology, and the division of the data sets into OD pair components. Then, in section 3, an analysis is performed for the four representative OD pairs, to find out whether a Gaussian IID assumption is a suitable approach for modelling. We conclude that at this level of OD pair resolution, where we have low level of aggregation, the model is quite unsuitable in many aspects. Finally, in section 4 we study the validity of a functional relationship between the mean and variance of an OD pair. The results indicate that such a relation exists, but for some periods of time it is rather weak.

# 2 Funet data

## 2.1 Measurement methodology

Traces were captured by Endance DAG 4.23 cards from 2.5 Gbit/s STM-16 link connecting nodes csc0-rtr and helsinki0-rtr in Funet network[1]. The IP addresses on captured packet headers were anonymized preserving prefix, and the headers were stored to disk using flow-based compression [9]. Captured traffic was transferred once an hour to an analysis machine, where statistics were calculated. Part of the traces were archived for later analysis.

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

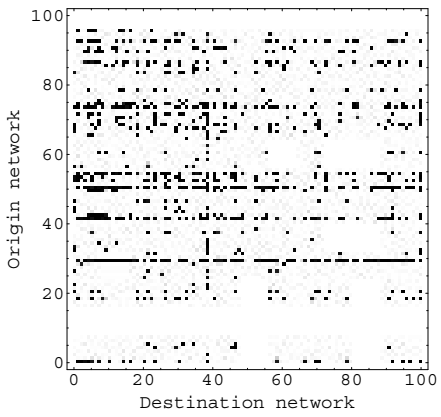


**Figure 1: Traffic matrix. Black indicates active OD pair, white inactive.**

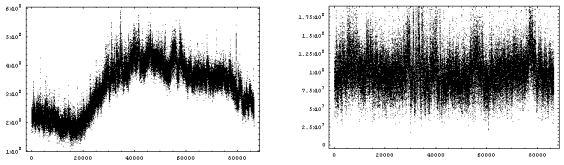

**Figure 2: One day traffic trace (bits/s) from the studied link in Funet network. Total traffic is on the left side, and aggregate of the studied OD pairs on the right side.**

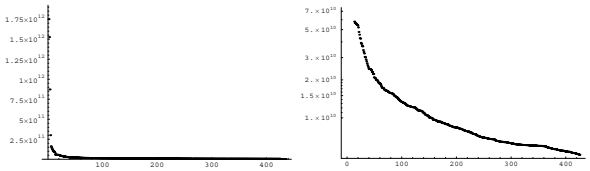

**Figure 3: The magnitudes (bits) of the OD pairs in sorted order at linear (left hand side) and logarithmic scale (right hand side).**

TCP accounted for more than 98 % of bytes transferred. During daytime 10-20 % of TCP traffic was HTTP (TCP port 80). There existed also considerable amount of peer-to-peer traffic.

The traffic was divided into origin-destination pairs by classifying in terms of IP addresses. We used a 22 bit network mask, meaning that each origin subnetwork could have 10 bits for host part or about 1000 IP addresses. This corresponds to a middle size company. As the resolution here is quite high, there are $2^{22}$, or over four million, origin networks. By selecting 100 largest networks based on traffic sent, we obtain a traffic matrix with 10000 origin-destination pairs. In practice, only 844 of these 10000 OD pairs had traffic on the measured link. This is depicted visually in Figure 1 which demonstrates the traffic matrix of 100 largest network with black boxes denoting active OD pairs.

## 2.2 Original data and its derivatives

The measurements considered in this paper capture the traffic of one day, November 30th 2004. We denote this original measurement data by $\boldsymbol{x} = (x_{t,k}; t = 1, 2, \ldots, T, k = 1, 2, \ldots K)$, where $x_{t,k}$ refers to the measured bit count of OD pair $k$ over one second period at time $t$ seconds. For each time scale $\Delta$, we created the corresponding time series of OD pair bit counts $x_k^\Delta = (x_{n,k}^\Delta; n = 1, 2, \ldots, T/\Delta)$ by defining

$$x_{n,k}^\Delta = \frac{1}{\Delta} \sum_{t=n\Delta+1}^{(n+1)\Delta} x_{t,k}.$$

All traffic using the studied link is shown on the left side of Figure 2, at one second resolution. The diurnal variation of the traffic at this level of aggregation is visible. On the right side of the figure we have plotted the aggregated traffic of the studied OD pairs, which comprises $33\%$ of the total link traffic. Surprisingly, no diurnal pattern is visible in this figure.

For OD pair $k$ we define the *magnitude* $X_k$ as the total number of bits transferred in one day,

$$X_k = \sum_{t=1}^{T} x_{t,k}.$$

The magnitudes of the OD pairs in descending order are shown in Figure 3 at linear and logarithmic scale. The magnitude decreases dramatically after the ten largest OD pairs.

Next we consider traffic of individual OD pairs. As in [7], we split the OD pair bit counts $x_{n,k}^\Delta$ into components,

$$x_{n,k}^\Delta = m_{n,k}^\Delta + s_{n,k}^\Delta z_{n,k}^\Delta,$$

where $m_{n,k}^\Delta$ refers to the moving sample average, $s_{n,k}^\Delta$ to the moving sample standard deviation, and $z_{n,k}^\Delta$ to the sample standardized residual of OD pair $k$. Based on Figure 2, the averaging period was chosen to be (about) 1 hour. Thus,

$$m_{n,k}^\Delta = \frac{1}{3600/\Delta + 1} \sum_{j=n-1800/\Delta}^{n+1800/\Delta} x_{j,k}^\Delta$$

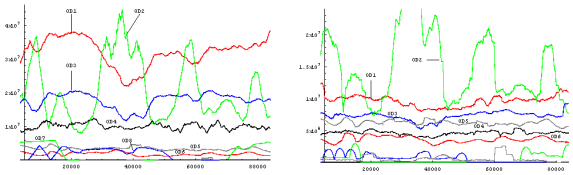

**Figure 4: Moving sample average (left) and sample standard deviation (right) for bit counts of largest OD pairs.**

and

$$s^{\Delta}_{n,k} = \sqrt{\frac{1}{3600/\Delta + 1} \sum_{j=n-1800/\Delta}^{n+1800/\Delta} (x^{\Delta}_{j,k} - m^{\Delta}_{j,k})^2}$$

The moving average, $m^{\Delta}_{n,k}$, is depicted for biggest OD pairs as a function of time on the left side of Figure 4. The moving sample-standard-deviation for the same pairs is depicted on the right side of Figure 4. The order of the greatest OD pairs in terms of the moving average and sample-standard-deviation remains over one day. The only exception is the second greatest OD pair, which seems to be very bursty.

## 2.3 OD pair grouping

From the traces it is evident that the OD pairs behave differently, and it is hardly possible to model them with a common traffic model and distribution. In this section we try to find some common characteristics of the OD pairs and group them accordingly.

First, some of the OD pairs seem to approximately follow the Gaussian distribution. The largest OD pair in terms of total magnitude, the bit counts of which are depicted in the upmost row of Figure 5, is a good example of this. We will call this OD pair *"Normal"*. The original trace is on the left side of the figure. It has a long-term variation, which, however, does not follow the ordinary diurnal pattern such as in left side of Figure 2. Traffic of the OD pair is instead at its highest in the night and lowest in the day of Finnish time. On the right of Figure 5 is shown the residual component $z^{\Delta}_{n,k}$ of the trace.

Second, there are some very bursty OD pairs. During the busy periods traffic load is many times higher and variable than during other periods. Example of this behavior is the second largest OD pair depicted in the second row of Figure 5. We will call this OD pair *"Bursty"*. Note that there is a significant number of idle seconds in the original trace. For this reason also the residual component has a clear lower bound.

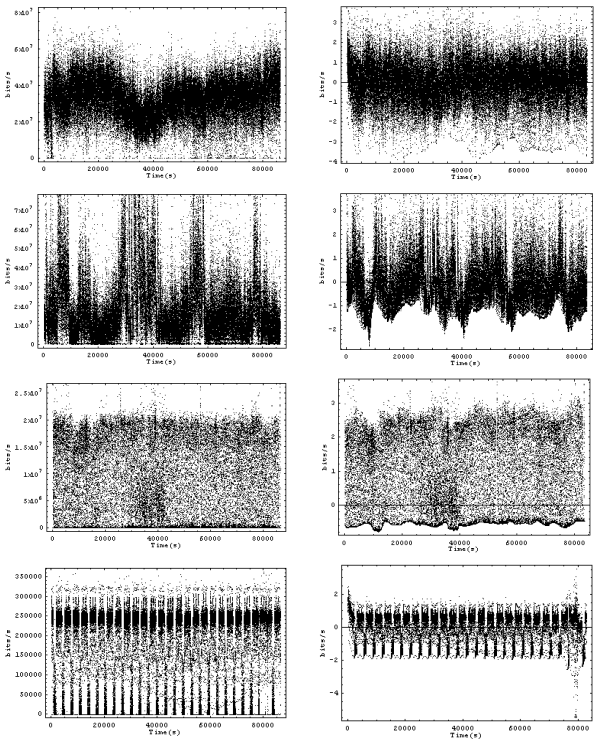

**Figure 5: Original traffic trace (bits/s) in the left side and residual component $z^{\Delta}_{n,k}$ in the right side of figure.From top to bottom: *Normal*, *Bursty*, *Uniform* and *Periodic* OD pair.**

Thirdly, some OD pairs seem to follow approximately an uniform distribution between zero and peak rate. This type of OD pair, depicted in the third row of Figure 5, we will call *"Uniform"*. Finally, some OD pairs are clearly periodic, as the last OD pair in Figure 5, which we will call *"Periodic"*.

At this very high level of address resolution none of the OD pairs seem to have a diurnal pattern. Even the largest OD pair (OD1 in Figure 4) that accounts roughly 10 percent of total traffic on the link, does not follow the diurnal pattern of the total link count. Also, as seen in Figure 2, the diurnal variation is next to non-existent in the combined traffic volumes of all the studied OD pairs. This would indicate that the strong diurnal variation in the link count is explained by a large number of small on-off OD pairs, that have long idle periods at night time. The diurnal pattern in the total link count arises from the changes in the number of OD pairs active at any given time, rather than a corresponding diurnal pattern in each OD pair.

The OD pair groups and the ranking of their representatives in terms of the total magnitude are listed in

**Table 1: The OD pair grouping and ranking in terms of the total magnitude of their representatives.**

| | Ranking | Name |
|---|---|---|
| 1 | 1 | *Normal* |
| 2 | 2 | *Bursty* |
| 3 | 5 | *Uniform* |
| 4 | 9 | *Periodic* |

Table 1. In next sections we will mainly analyze these four aforementioned representative of OD pairs.

# 3 Gaussian IID model

In this section we study the statistical properties of the OD pairs distinctly based on the grouping presented in the previous section. We examine the validity of the common assumption that OD pair traffic follows Gaussian distribution and both consecutive bit counts of a given OD pair and bit counts over two different OD pairs are independent.

## 3.1 Gaussian assumption

In this section we study whether the OD pairs follow the Gaussian distribution, as is often assumed in traffic engineering. We concentrate on the stochastic component, the standardize residual $z_n^{\Delta}$, as defined in section 2, and study measurements of one second and one minute intervals. One way to evaluate the appropriateness of the Gaussian assumption is the Normal quantile (N-Q) plot. The original sample $x$ is ordered from the smallest to the largest and plotted against $a$, which is defined as

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

where $\Phi$ is the cumulative distribution function of the Gaussian distribution. The vector $a$ contains the quantiles of the standard Gaussian distribution, thus ranging approximately from $-3$ to 3. If the considered data does indeed follow the Gaussian distribution, this 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.

$$r(x,a) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(a_i - \overline{a})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(a_i - \overline{a})^2}}. \quad (4)$$

The marginal distributions and N-Q plots of the four representative OD pairs are depicted in Figure 6 for one second aggregation level, and in Figure 7 for one minute aggregation level. The goodness of fits for Gaussian distribution of all four representative OD pairs are collected to Table 2.

**Table 2: Goodness of fit $r^2$ with regard to the Gaussian distribution**

| OD pair | $r^2$, $\Delta = 1$ | $r^2$, $\Delta = 60$ |
|---|---|---|
| Normal | 0.996 | 0.996 |
| Bursty | 0.936 | 0.976 |
| Uniform | 0.661 | 0.965 |
| Periodic | 0.816 | 0.698 |

On the upmost row of the figures, the histogram and N-Q plot of "Normal" OD pair show that the data trace seems to follow Gaussian distribution quite closely. The goodness of the fit for the N-Q plot is $r^2 = 0.996$ which indicates a reasonably good fit.

For the "Bursty" OD pair, depicted in the second row of Figures 6 and 7, the traffic does not follow Gaussian distribution. The marginal distribution of the trace indicates the OD pair is more heavy-tailed. Modelling the OD pair with log-normal distribution, for example, would be more reasonable.

The distribution of the "Periodic" OD pair is very far from Gaussian. The same holds for the "Uniform" OD pair at one second measurement interval. However, an interesting property for the "Uniform" OD pair is that the distribution looks reasonably close to Gaussian distribution with the 60 second measurement interval.

## 3.2 Dependence within OD pairs

In this section we study whether we can treat consecutive measurement samples of an OD pair as independent from each other, with regard of the stochastic component. If the traffic of an OD pair was to be IID, there should not be any significant autocorrelation in the residual component $z_{n,k}^{\Delta}$. The autocorrelation function for the residual is defined as:

$$r_l(k) = \frac{\sum_{i=1}^{T/\Delta - l}(z_{i,k}^{\Delta} - \bar{z}_k^{\Delta})(z_{i+l,k}^{\Delta} - \bar{z}_k^{\Delta})}{\sum_{i=1}^{T/\Delta}(z_{i,k}^{\Delta} - \bar{z}_k^{\Delta})^2},$$

where $T/\Delta$ is the size of time series and $l$ is the *lag* .

The autocorrelation functions (acf) for residuals of the four representative OD pairs are depicted in Figure 8. On the left side of the figure the aggregation level is one second ($\Delta = 1$) and on the right side it is one minute ($\Delta = 60$).

For the "Normal" and "Bursty" OD pairs, there are statistically significant autocorrelation values for the lags of several seconds, as depicted on the left side of the

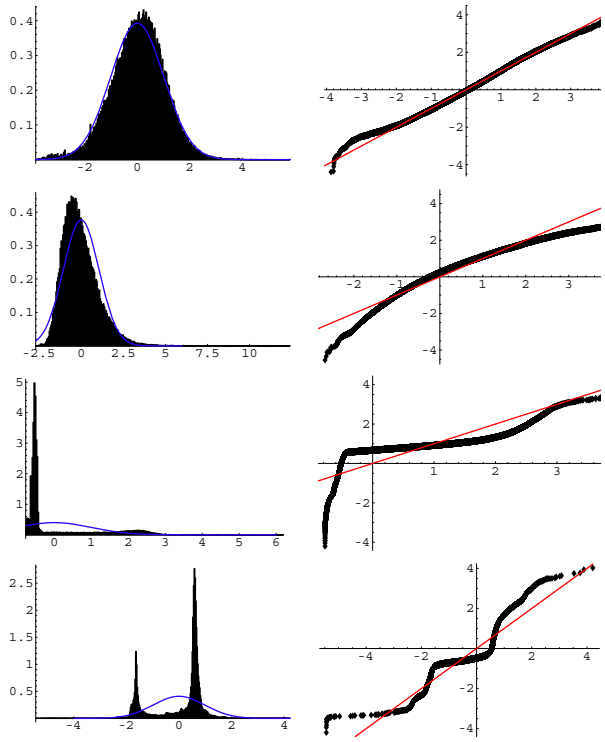

**Figure 6: Marginal distributions and N-Q plots for residuals with $\Delta = 1s$. From top to bottom: *Normal*, *Bursty*, *Uniform* and *Periodic* OD pair.**

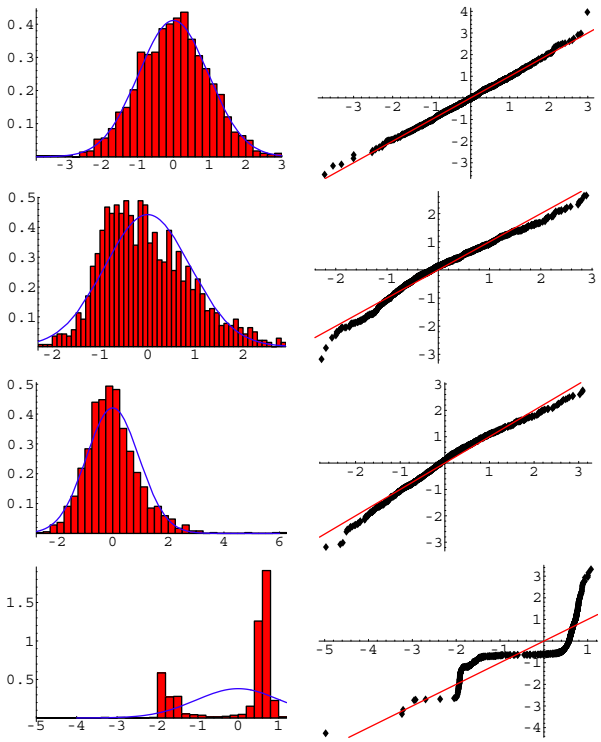

**Figure 7: Marginal distributions and N-Q plots for residuals, $\Delta = 60s$. From top to bottom: *Normal*, *Bursty*, *Uniform* and *Periodic* OD pair.**

figures. For longer lags, depicted on the right, the autocorrelation tends to zero, and is under the confidence interval shown with dashed lines, indicating that IID assumption probably is valid for that type of OD pairs. The exception is that for the "Normal" OD pair in the one minute aggregate there exist spikes at every 10 minutes.

For the "Uniform" OD pair the autocorrelation is almost zero even with very short lags of a few seconds.

The last row of Figure 8 depicts the autocorrelation function of the "Periodic" OD pair. This is the only OD pair which clearly has long range dependence, in this case because of the periodicity. Indeed the periods can be clearly observed from the correlation structure.

It is worth of noting that for the "Normal" OD pair, we can see strong spikes approximately 5 seconds apart from each other. A closer inspection of the trace revealed that traffic decreased dramatically every 5 seconds. Also the "Bursty" OD-pair has spikes in its autocorrelation function every 5 seconds, but these are lot smaller than in the "Normal" OD pair. The reason for this behavior is most likely in the origin network that the two representative OD pairs share.

### 3.3 Dependence between OD pairs

In this section we study whether the OD pairs are independent from each other. We study the dependency between the residual components of the OD pairs, because a possible long term traffic variation originating from similar diurnal behavior would increase the correlation between the OD pairs.

To evaluate the dependency between OD pairs we have calculated cross-correlation between the residuals $z^{\Delta}_{n,k}$ and $z^{\Delta}_{n,k'}$ of different OD pairs $k$ and $k'$:

$$r(k,k') = \frac{\sum_{i=1}^{n}(z^{\Delta}_{i,k} - \bar{z}^{\Delta}_{k})(z^{\Delta}_{i,k'} - \bar{z}^{\Delta}_{k'})}{\sqrt{\sum_{i=1}^{n}(z^{\Delta}_{i,k} - \bar{z}^{\Delta}_{k})^2 (z^{\Delta}_{i,k'} - \bar{z}^{\Delta}_{k'})^2}}.$$

For that purpose we have selected 20 OD pairs from the traffic matrix in Figure 1. The correlation values are presented graphically in Figure 9. In the figure we have not considered the correlation between a given OD pair and itself, which would obviously equal 1.0. The distribution of the various correlation terms is also shown in the same Figure 9. The horizontal lines in the figure depict the 95% confidence interval of the hypothesis

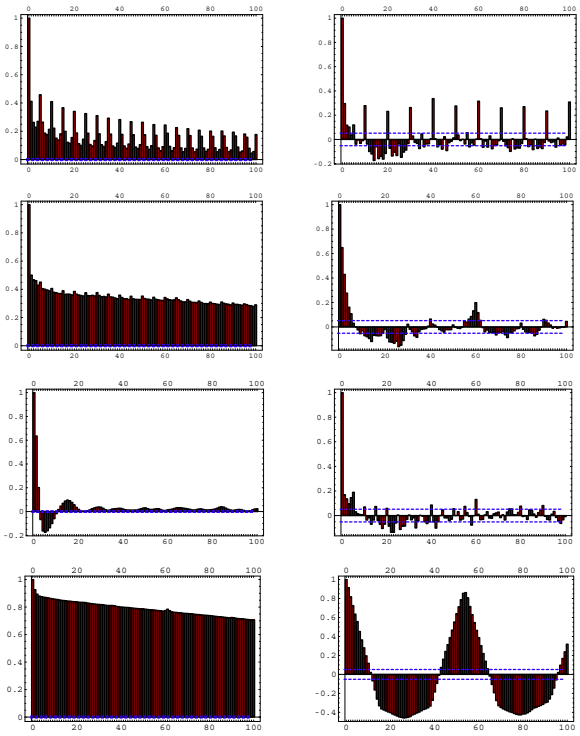

**Figure 8: Autocorrelation function for residual component. Left side: 1 second resolution, right side: 60 second resolution. From top to bottom: *Normal*, *Bursty*, *Uniform* and *Periodic* OD pair.**

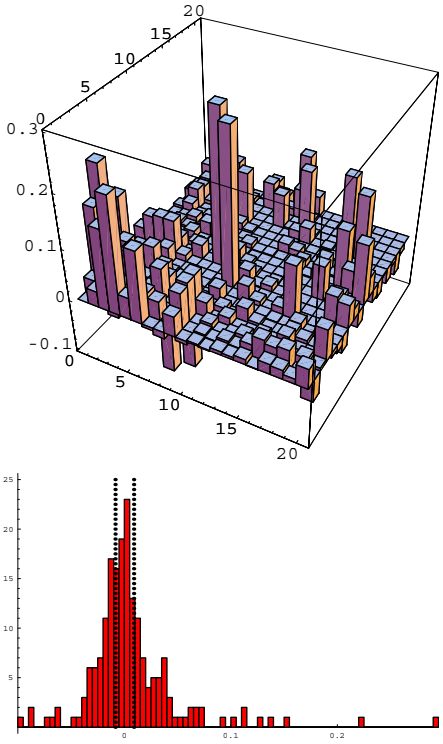

**Figure 9: Top: Correlation between 20 greatest OD pairs, bottom: distribution of correlation coefficients, with 95% confidence interval depicted by dotted lines.**

that correlation would be zero. Clearly there are a large number of statistically significant non-zero values in our data.

To better understand the correlation between OD pairs, we concentrate on those pairs of OD pairs that have the greatest correlation, either positive or negative. Table 3 lists 10 such pairs together with the origin and destination networks. As we can see, it is not only the pairs that share a common origin (or destination) network that are correlated. Also the pairs of OD pairs that have completely different origin and destination networks can have significant correlation between them.

**Table 3: Origin and destination networks of pairs of OD pairs and cross-correlations between them.**

| $s_1$ | $d_1$ | $s_2$ | $d_2$ | $r$ |
|---|---|---|---|---|
| 65 | 42 | 51 | 36 | 0.29 |
| 1 | 5 | 1 | 2 | 0.22 |
| 1 | 4 | 1 | 2 | 0.15 |
| 1 | 21 | 1 | 2 | 0.14 |
| 30 | 5 | 66 | 5 | 0.13 |
| 30 | 46 | 30 | 4 | 0.11 |
| 30 | 5 | 34 | 1 | 0.11 |
| 30 | 5 | 23 | 35 | 0.10 |
| 66 | 5 | 23 | 35 | 0.10 |
| 51 | 36 | 1 | 3 | -0.10 |

# 4 Mean-variance relationship

In this section we study the mean-variance relation. We separate between two different situations that are not to be confused with one another. First, by *temporal relation* we mean that the variance of a particular OD pair's traffic at a given time is related to the volume of the traffic at that time. That is, when there are more traffic, also the variation is higher. The second situation is the *spatial relation* in which we consider the relation over OD pairs or links. That is, it is studied whether the variance of an OD pair is larger for the OD pairs that have larger traffic volumes. This is a key assumption in many traffic matrix estimation techniques.

## 4.1 Temporal relation

Regarding the temporal relation, Medina et al. [1] observed that the exponent $c$ varies remarkably from one OD pair to another within bounds $c \in [0.5, 4.0]$. These observations were based on data collected from a Tier-1 backbone with measurement period $\Delta = 1$ s. Soule et al. [6] found similar results from flow data collected from a commercial Tier-1 backbone with measurement period $\Delta = 300$ s. They report that the power-law constant $c$ for individual OD flows varies in a range spanning from 1 to 4. In our previous work [8] we found that this is true also for the Lucent local area network, used in [7]. For some OD pairs the fit was reasonably good, but for others it was next to non-existent.

In this paper we study the temporal mean-variance relation for the four representative OD pairs defined in previous sections. In the results presented here, the mean and variance are calculated for each one hour period of the 24 hour trace, but using different intervals yields similar results. The mean and variance values for a given hour comprises one point in the plots of Figure 10 and 11, where mean is depicted on the horizontal axis, and variance on the vertical axis. The figures thus have 24 points presented in log-log scale. The logarithm of the mean-variance power law relation (3) is

$$\log D^2[X_{n,k}] = c \log E[X_{n,k}] + \log \phi.$$

Thus, in log-log scale the exponent $c$ is a linear coefficient. If the relation would hold, the points would fall on a line with slope $c$ and intercept $\log \phi$.

Figure 10 depicts the situation when measurement interval is one second, and Figure 11 when it is 60 seconds. The best linear fit in the least square sense is depicted in the figures. The only OD pair out of the four

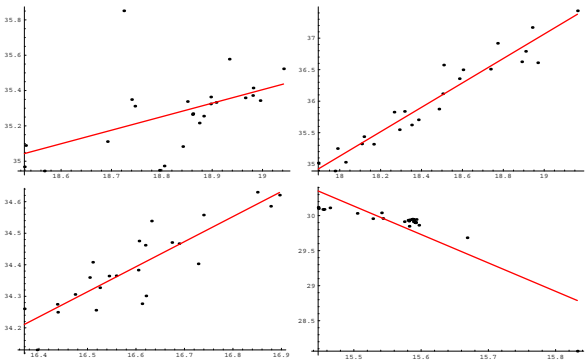

**Figure 10: Temporal mean-variance relation on one second resolution: top row: *Normal* and *Bursty* OD pairs, bottom row *Uniform* and *Periodic* OD pairs.**

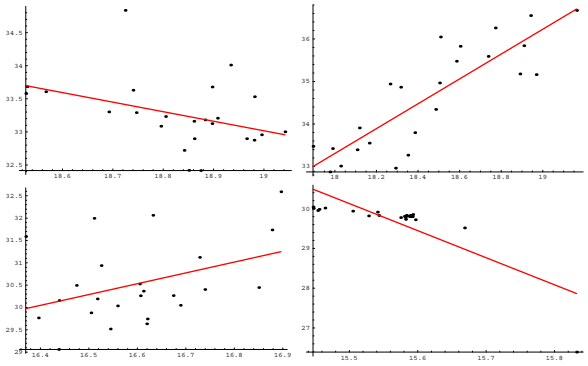

**Figure 11: Temporal mean-variance relation on 60 second resolution. Top row: *Normal* and *Bursty* OD pairs, bottom row: *Uniform* and *Periodic* OD pairs.**

that portrays behavior that would indicate the existence of a mean-variance relation is the Bursty OD pair, which has a goodness of fit value of $R^2 = 0.92$ for the relation, and estimated parameter value of $c = 1.9$. For the other OD pairs the fits are much worse. Changing the measurement intervals made the fit worse for the Bursty OD pair, with $R^2 = 0.76$ and $c = 2.9$ now. For the other pairs the changed measurement interval changes the $c$-parameter even more dramatically. This is due to the fact that when there is not much of a relation, some parameter value still always gives the best fit, and that value may change significantly, even due to a small change in the data.

In general, there does not seem to be a temporal mean-variance relationship in our data.

## 4.2 Spatial relation

In Cao et al. [7] the authors consider only integer values for the spatial mean variance relation and conclude that $c = 2$ gives reasonably good fit. Our study [8] of the same local area network data set yielded parameter value of $c = 1.96$. Medina et al. [1] reported that the mean-variance relationship seems to hold, with the parameter value being $c = 1.97$ in their study. Gunnar et al. [2] also confirmed the validity of the spatial relation, and reported parameter values $c = 1.5$ and $c = 1.6$ based on data traces from a global operator's backbones in US and Europe, respectively. Soule et al. [6] found that in a backbone network studied, the best linear fit resulted in value $c = 1.56$, but did not find the fit sufficiently good to justify the use of the relation. However, they remain uncertain as to the effect this inaccuracy brings to the estimation results. We will study this in section 4.2.1 through a simulation study. Also, Soule et al. express

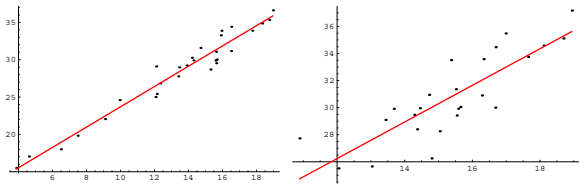

**Figure 12: Spatial mean-variance relation, examples of good and not so good fits. Left: $R^2 = 0.95$, right: $R^2 = 0.70$**

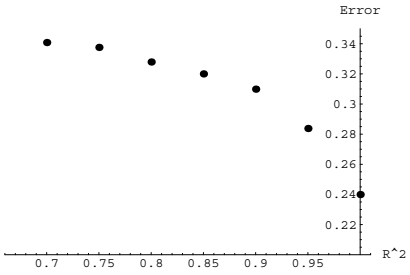


**Figure 13: Errors in traffic matrix estimation as a function of the goodness of fit of the mean-variance relation.**

concerns that the varying parameter values of temporal relation might hinder the use of the spatial relation in traffic matrix estimation.

We study the spatial relation between the 40 largest OD pairs in the Funet data, concentrating on one hour periods that we assume to be approximately locally stationary. The average goodness of fit is $R^2 = 0.83$ with the values ranging from 0.60 to 0.95, while the estimate for the exponent parameter varies from 1.11 to 1.46, with 1.34 being the average. For some hours the fit is reasonably good, as shown in Figure 12, where one point in the plot depicts the mean and variance of one OD pair during the studied hour. For some other one-hour periods the fit is not very good. Changing the measurement interval to 60 seconds or taking more OD pairs into consideration does not affect the situation significantly.

We can conclude that for our OD pairs, there is a vague spatial mean-variance relation, with good fits for some one hour periods. It is to be noted however, that we have an extremely high resolution in dividing the trace into these virtual OD pairs, so the situation is different from the typical traffic matrix estimation situation, where the OD pairs are larger and thus more aggregated, as in the study by Gunnar et al. for instance. This might affect also the validity of the mean-variance relation.

#### 4.2.1 Effects of inaccuracies in the spatial relation

An important aspect regarding what we can conclude about the validity of the mean-variance relation is to study the actual effect that inaccuracies cause in traffic matrix estimation, where we try to find an estimate for the OD pair traffic volumes, based on the link count measurements.

We performed a simulation study for this purpose, using a six node topology with 30 OD pairs. Synthetic data sets were created, where the mean-variance relationship holds to different degrees. For each goodness of fit value we performed the simulation several times by drawing new set of synthetic Gaussian measurements of sample size 100, with the underlying parameters staying the same. After obtaining maximum likelihood estimates for the traffic matrix by the EM algorithm (see [7]), the average error of the estimates is then computed for each scenario.

The results of the errors as a function of the goodness of fit value for the mean-variance relation used in the simulations can be seen in Figure 13. The effect of a bad fit is not as dramatic as one might think. Even with $R^2 = 0.70$ the errors are less than 1.5 times larger than in the ideal situation, This is due to the fact that maximum likelihood estimates are dominated by the first order equation, and the mean-variance relation is used only to get the second order terms to bring in the extra information needed to make the ill-posed problem identifiable.

The reasonable accuracy is all the more surprising, when remembering that the $R^2$-values are computed in log-log scale. Although a 0.70 goodness of fit, as depicted in Figure 12, looks reasonably good in that scale, in linear scale the accuracy is not very good.

However, we must note that at values close to 1.0 the situation is quickly deteriorating as the fit becomes worse. In our data set the average fit was $R^2 = 0.83$. Around that kind of values, a change of 0.05 in the goodness of fit to either direction is not too critical, but there is a clear price to be paid in estimation accuracy for the fact that the relation does not hold exactly.

## 5 Conclusion

In this paper we have studied the characteristics of OD pair traffic, by dividing the traffic in a Funet link into OD pair components based on source and destination addresses on the captured packets. The division was made with 22 bit mask, so the level of aggregation is low in our OD pairs, as one origin node corresponds to only 1000 IP addresses. With this resolution of aggregation

we found that many of the OD pairs are quite far from Gaussian distribution, as opposed to the aggregated traffic on the link which follows Gaussian distribution rather closely. We identified four typical OD pairs, and studied these representative OD pairs in further detail.

The "Normal" OD pair is the only one that is clearly Gaussian. It is in many regards close to the behavior of the aggregated link traffic, except for the diurnal pattern. In fact, we noticed that none of the OD pairs had the kind of daily variation prevailing in the aggregate. Thus, the cause of the diurnal pattern must be smaller on-off flows, that are idle during the night.

The "Bursty" OD pair has a wider range of variation, but is still somewhat Gaussian with the longer $60$ second measurement interval. Both "Bursty" and "Normal" OD pairs have autocorrelation similar to that of the overall aggregation, where there are five minutes of positive values and then a set of negative values around $20$ minute lag. The "Bursty" OD pair is also the only one of the representative OD pairs that follows a temporal mean variance relation, such that higher traffic volumes mean also higher variation.

The "Uniform" OD pair has an uniform distribution between zero and peak rate when it is not idle. With the longer $60$ second aggregates it is, however, much closer to Gaussian.

In the "Periodic" OD pair the periodicity is clear even with the longer aggregates, and seen also clearly in the behavior of the autocorrelation function.

As a result, the Gaussian IID model is not generally applicable for these OD pairs. While many of the $60$ second aggregates are at least somewhat Gaussian, this is not true for all OD pairs. Also there are significant non-zero autocorrelation values. And finally, the OD pairs are not independent, but we found cross-correlation values up to $0.30$, and large number of smaller yet still statistically significant non-zero values.

Finally, there is a vague spatial mean-variance relation, with exponent parameter $c = 1.34$, meaning that the OD pairs with larger means do also have larger variance according to a power law relation. Over some one-hour periods this relation is quite accurate, but at other times only somewhat indicative. However, we illustrated that this inaccuracy does not affect the accuracy of the maximum likelihood estimates in traffic matrix estimation as dramatically as one might suspect. If the relation is not spot-on, which it rarely is, the decline in the accuracy is not that big when we go from reasonably good fit to moderate fit.

Further work could include shortening the address classification part in creating the virtual OD pairs, thus obtaining smaller number of pairs, with more aggregated traffic. It would be interesting to see at which level of aggregation the diurnal pattern becomes visible in OD pair traffic.

## Acknowledgment

The authors would like to thank CSC - the Finnish IT center for science - for providing access to Funet network and for computing and archive resources.

## References

[1] A. Medina, N. Taft, K. Salamatian, S. Bhattacharyya, and C. Diot, Traffic matrix estimation: Existing techniques and new solutions, in SIGCOMM'02, Pittsburg, USA, August 2002.

[2] A. Gunnar, M. Johansson, and T. Telkamp, Traffic matrix estimation on a large IP backbone – A comparison on real data, in IMC'04, Taormina, Italy, October 2004.

[3] S. Vaton, J.S. Bedo, A. Gravey, "Advanced methods for the estimation of the Origin Destination traffic matrix", Revue du 25me anniversaire du GERAD, 2005.

[4] A. Soule, A. Lakhina, N. Taft, K. Papagiannaki, K. Salamatian, A. Nucci, M. Crovella, and C. Diot, Traffic matrices: Balancing measurements, inference and modeling, in SIGMETRICS'05, Banff, Canada, June 2005.

[5] A. Lakhina, K. Papagiannaki, M. Crovella, C. Diot, E.D. Kolaczyk, and N. Taft, Structural analysis of network traffic flows, in SIGMETRICS/Performance'04, New York, USA, June 2004.

[6] A. Soule, A. Nucci, R. Cruz, E. Leonardi, and N. Taft, How to identify and estimate the largest traffic matrix elements in a dynamic environment, in SIGMETRICS/Performance'04, New York, USA, June 2004.

[7] 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.

[8] I. Juva, R. Susitaival, M. Peuhkuri, and S. Aalto, Traffic characterization for traffic engineering purposes: Analysis of Funet data, NGI 2005, Rome, Italy, April 2005.

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