Ilmari Juva, Riikka Susitaival, Markus Peuhkuri, and Samuli Aalto. Effects of spatial aggregation on the characteristics of origin-destination pair traffic in Funet. Helsinki University of Technology, Networking Laboratory, Report 1/2007, ISBN 978-951-22-8681-2, 2007.

Effects of Spatial Aggregation on the Characteristics of Origin-Destination Pair Traffic in Funet

Ilmari Juva, Riikka Susitaival, Markus Peuhkuri, and Samuli Aalto
TKK Helsinki University of Technology, Networking Laboratory
ilmari.juva@tkk.fi

March 2, 2007

Abstract

In this paper we analyze measurements from the Finnish University Network (Funet) and study the effect of spatial aggregation on the origin-destination flows. The traffic is divided into OD pairs based on IP addresses, using different prefix lengths to obtain data sets with various aggregation levels. We find that typically the diurnal pattern of the total traffic is followed more closely by the OD pairs as their volume increases, but there are many exceptions. Gaussian assumption holds well for all OD pairs when the aggregation level is high enough, and we find an approximate threshold for OD pair traffic volume after which they tend to be Gaussian. Also the functional mean-variance relation holds better when the aggregation level is higher.

1 Introduction

Origin-Destination (OD) pair traffic refers to the traffic flow that traverses between two nodes in a network. Depending on aggregation level, these can be, for example, hosts, routers, or ISPs. The main feature of measuring OD pair traffic is that traffic has to be aggregated both in time and space. Diurnal variation of the Internet traffic is usually studied at the coarse level of temporal aggregation with sample interval of some minutes whereas the packet level dynamics has to be studied at a very fine granularity of time. Traffic flowing between two hosts is example of very fine level of spatial aggregation, whereas ISP level studies is example of coarse aggregation in space.

In many areas of traffic engineering, nature of OD pair traffic plays an important role. For example, in traffic matrix estimation one estimates the OD traffic flows from the measured link loads. The existing estimation techniques make several assumptions about the OD pair traffic, including Gaussianity, functional mean-variance relationship and independence of the traffic samples. Clearly, the validity of these assumptions in real traffic traces depends both on level of temporal and spatial aggregation.

Few papers have studied the characteristics of OD pair traffic earlier. First, Feldman et al. [3] characterize point-to-multipoint traffic and find that a few demands account for 80% of total traffic and the traffic volumes follow Zipf's law. Daily profiles of the greatest demands also vary significantly from each other. Bhattacharyya et al. characterize Point of Presence-level (POP) and access-link level traffic dynamics in [4]. Also they find that there are huge differences in the traffic volumes of the demands. In addition, the larger the traffic volume of an egress node, the larger also the variability of traffic during the day. Finally, Lakhina et al. [5] analyze traffic of two backbone networks. Using Principal Component Analysis (PCA) they demonstrate that OD flows can be 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, spiky and noisy. We have also previously studied the characteristics of traffic from Funet network link measurements . In [7] we studied the characteristics of

aggregate link traffic and in [8] OD pair traffic at a fixed spatial aggregation level. Even though these aforementioned measurement studies answer to some questions related to OD pair traffic, full understanding how spatial aggregation changes the characteristics of OD pair traffic, is still missing.

To this end, in this paper we study the effect that aggregation in space has on the OD pair traffic characteristics. The traffic of the link in Funet network is divided into OD pairs with different prefix lengths. Often traffic characteristics are analyzed in short time scales. We take the vantage point of traffic engineering and traffic matrix estimation, in which the relevant time scale is minutes, instead of seconds or less. We show that while the diurnal pattern of the OD pairs is not always the same as the diurnal pattern of the total traffic, the correlation is better, in general, as the OD pair's traffic volume is larger. The Gaussian assumption, on the other hand, is shown to hold well for all OD pairs over a certain size. For the relation between mean and variance we found that the larger the aggregation level, the better the relation holds.

The rest of the paper is organized as follows. In Section 2 we explain the measurement methodology and introduce the data set used in the study. Section 3 studies the magnitudes of OD pairs, while Sections 4 and 5 study how the aggregation affects the diurnal pattern and gaussianity of the OD pairs. In Section 6 the existence of a mean-variance relation is studied. Finally, section 7 concludes the paper.

2 Measurements and original data

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 link is two-directional and we denote the direction from helsinki0-rtr to csc0-rtr by d_0 and the opposite direction by d_1. Further details of the measurement process are available in earlier work based on the same measurements [7].

We divide the traffic of the link into origin-destination pairs by identifying the origin and destination networks of packets by the left-most bits in the IP address. Let l denote the number of bits in this network prefix, also called network mask. Different levels of aggregation are obtained by changing the prefix length l. The maximum length of the network prefix is 24 bits. With this resolution, there are 2^{24}, or over sixteen million, possible origin networks. On the other hand, with the prefix length $l = 1$ there are only two networks and thus four possible OD pairs.

Our procedure for selecting OD-pairs for further analysis from the original link traffic is the following. Combining both directions, the N most active networks in terms of traffic sent are selected and a $N \times N$ traffic matrix is formed, where $N \leq 100$. From the obtained traffic matrix at most M greatest OD pairs in terms of sent traffic are selected for further analysis. We select $M = 100$, except in section 6, where we use $M = 1000$. Note that for very coarse level of aggregation the number of all OD pairs remains under 100.

The measurements capture the traffic of two days: November 30th 2004 and June 31st 2006, with the main focus being on the first day. The traffic is divided into origin-destination pairs using different prefix lengths and aggregated in time to one minute resolution. For each prefix length l and direction d_0/d_1 separately, we denote the original measurement data by

$$\mathbf{x} = (x_{t,k}; t = 1, 2, \ldots, T, k = 1, 2, \ldots, K),$$

where $x_{t,k}$ refers to the measured bit count of OD pair k over one minute period at time t minutes.

Let us consider traffic of individual OD pairs. As in [1], we split the OD pair bit counts $x_{t,k}$ into components,

$$x_{t,k} = m_{t,k} + s_{t,k} z_{t,k},$$

where $m_{t,k}$ refers to the moving sample average, $s_{t,k}$ to the moving sample standard deviation, and $z_{t,k}$ to the sample standardized residual of OD pair k. The averaging period was chosen to be one hour. Thus,

$$m_{t,k} = \frac{1}{60} \sum_{j=n-30+1}^{n+30} x_{j,k}$$

and

$$s_{t,k} = \sqrt{\frac{1}{60} \sum_{j=n-30+1}^{n+30} (x_{j,k} - m_{j,k})^2}$$

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

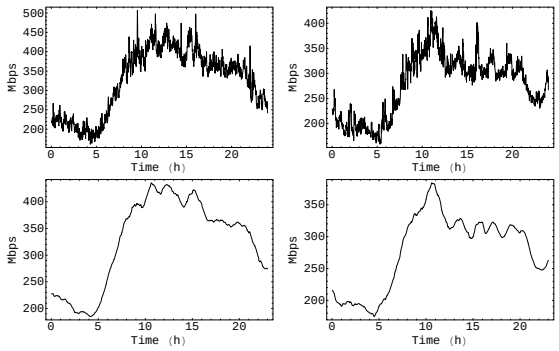

Figure 1: One day traffic trace of the studied link. Left side: direction d_0, right side: direction d_1.

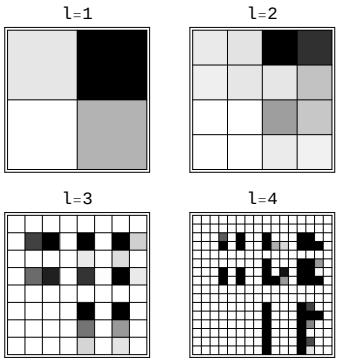

Figure 2: Traffic volume sent between the origin and destination network for different prefix lengths l. Black: a lot of traffic, white: no traffic. Direction d_0.

The traces of total traffic on the first measured day in the studied link for directions d_0 and d_1 are shown in the left and right side of Figure 1, respectively. The figure depicts also the moving sample averages of the traces. The diurnal variation of the traffic at this level of aggregation is clearly visible. The busiest hour of the day is in the middle of the day from 11 a.m. to 12 a.m. in both directions.

3 Magnitudes of OD pairs

In this section we study the size of the OD pairs at different aggregation levels. We are interested in how the traffic is distributed in address space, and whether there is a power law behavior observable in the sizes of the OD pairs, which would mean that the decrease in OD pair size as a function of rank should be linear in the log-log scale.

For OD pair k we define the *volume* X_k as the average of bits transferred per second over one day,

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

When the level of aggregation is very coarse ($l \leq 4$), the number of non-zero OD pairs is smaller than 100 and we are able to study the volumes of the complete traffic matrix. In Figure 2 we have depicted traffic matrices for cases from $l = 1$ to $l = 4$. In the density graphs the darker the color is the more traffic is sent, while white indicates that there is no traffic between the networks. When $l = 1$, the classification into OD-pairs is done based on the first bit of the network prefix. The density plot shows that most of the traffic in the link originates and terminates in the network whose first bit of prefix is 1. On the other hand, there is no traffic at all between networks with first bit 0. As we increase l, the density plots become sparser since the non-zero OD pairs form only a minor part of all possible OD pair combinations in the traffic matrix. One reason for sparseness is that the measurements are not network wide, but just from one link.

Next we consider the volumes of the OD pairs with

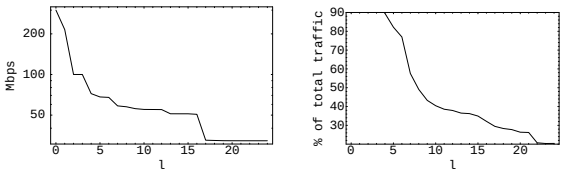

Figure 4: Left side: The volumes of the greatest OD pairs as a function of prefix length l. Right side: The percentage of traffic of 15 greatest OD pairs as a function of l. Direction d_0.

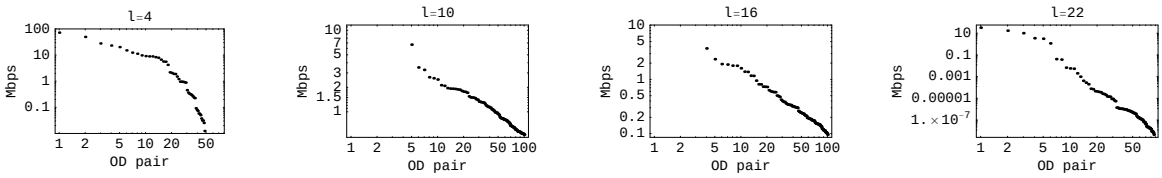

Figure 3: Traffic volume of OD pairs for different prefix lengths l. Direction d_0.

different values of l. In Figure 3 the OD pairs are sorted from greatest to smallest and their volumes are plotted on the log-log scale, when the prefix length varies from $l = 4$ to $l = 22$. For every level of aggregation there are approximately 15 very significant OD pairs and after that the volumes decrease. We note that for $l \geq 10$ the decrease is quite linear.

On the left side of Figure 4 the volume of the greatest OD pair for each aggregation level l is plotted. Decrease in the volume as a function of l is first very steep even in the logarithmic scale, but then it saturates until l changes from 16 to 17 where the volume drops again. In general, as compared to the hypothetical situation where all link traffic is divided evenly among all possible OD-pairs, the decrease is moderate. On the right side of Figure 4 we show the percentage that the 15 greatest OD pairs comprise of the total link traffic as a function of l. Even for finer resolutions, such as $l = 16$, these 15 pairs form a significant part of the traffic.

As a result of this section we can say that the classification of the link traffic based on origin and destination pairs produces "mice" and "elephants", which is a well known phenomenon from earlier Internet measurement studies. However, the power-law assumption is valid only for finer granularity of aggregation, such as $l \geq 10$, where the traffic volumes are smaller.

4 Diurnal variation of the OD pair traffic

In [8] we observed that at a fine aggregation level of $l = 22$ none of the OD pairs seemed to follow the diurnal variation of the total link traffic, in which the traffic peaks in the midday. We concluded that the strong diurnal variation in the link traffic is more explained by the variation in the number of active on-off OD pairs than diurnal pattern within these OD pairs. However, we would expect that when increasing the aggregation level, at some point the diurnal pattern should become visible in the OD pairs.

In this section we study in more detail the diurnal variation of the OD pairs at different levels of OD pair aggregation. This is done by comparing the daily profiles of the OD pairs and the corresponding profile of the total link traffic, shown in the lower row of Figure 1. As an example, we plot the moving sample averages of the four largest OD pairs with aggregation levels $l = 4$ and $l = 8$ for direction d_0 in Figure 5. At the coarse level of aggregation we can see different types of diurnal patterns. Pairs 3 and 4 have a diurnal variation close to the variation of the total link traffic, while pairs 1 and 2 are not so close. At the resolution $l = 8$ only the fourth OD pair follows the diurnal pattern of the link traffic.

To better understand how the diurnal variation changes as the aggregation level l increases, we study the correlation between two time series; the moving sample average of the total link traffic, and moving sample average of the OD pair k. The correlation coefficient between any two time series $x = (x_i, i = 1, ..., n)$ and $y = (y_i, i = 1, ..., n)$ is defined as

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

On the left side of Figure 6 we plot the correlation coefficients for all OD pairs with all aggregation levels l and directions d_0 and d_1 as a function of the volume of the OD pair. For small OD pairs there exists both positive and negative correlations but for large OD pairs the correlations are positive, as we would expect. However,

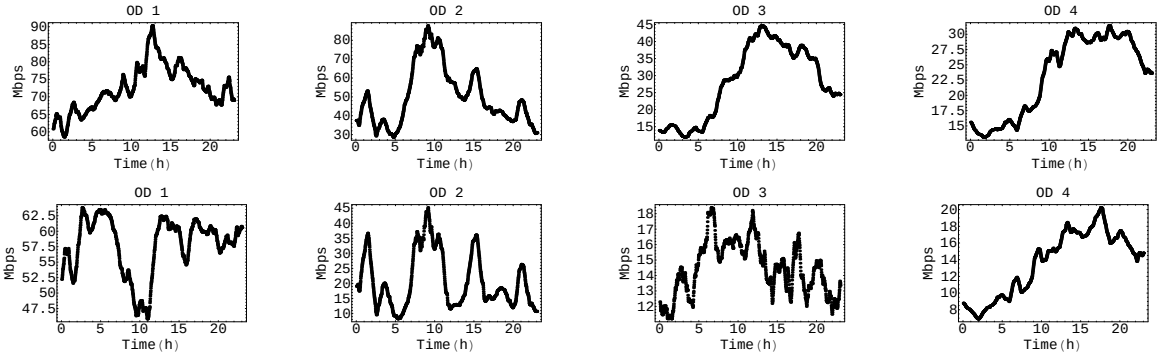

Figure 5: The moving sample average for the 4 greatest OD pairs. Prefix length $l = 4$ (upper) and $l = 8$ (lower). Direction d_0.

dependence between the correlation and the volume of the OD pair is not strong. In the right hand side of the same figure the mean of the correlation coefficients for the OD pairs with given prefix length l are plotted. We can see that the mean correlation decreases as a function of l, as the earlier figures indicated.

As a conclusion of this section we can state that as the aggregation level of the traffic coarse, also the diurnal traffic pattern of the OD pairs is closer to the variation of the total link traffic. However, there is not any clear bound in OD pair volume or in the prefix length, after which we can say that the daily behavior is similar to the familiar profile found in the link traces.

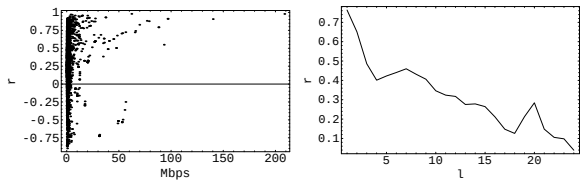

Figure 6: Testing diurnal variation. Right side: OD pairs correlation to the total link traffic as a function of the traffic volume. Left side: Average correlation of OD pairs with different prefix lengths l.

5 Gaussianity

In [7] the aggregated link traffic was found to follow very closely the Gaussian distribution. However, when we studied the origin-destination flows in [8], only a small portion of them were anywhere close to Gaussian, typically only the larger flows. Due to the Central Limit Theorem we might assume that when the aggregation of individual non-gaussian flows is large enough, the aggregate will indeed follow the Gaussian distribution. In [9] the authors studied the number of users required for aggregate to be Gaussian and found that "a few tens of users" is typically sufficient. We study the different aggregation levels in terms of traffic volume in order to determine how much traffic is needed to yield Gaussian behavior.

We evaluate the Gaussianity of each OD pair by the Normal-quantile (N-Q) plot of the standardized residual $z_{t,k}$. The original sample (denoted by x in the equation) 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 Φ 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 follows the Gaussian distribution, the N-Q plot should be

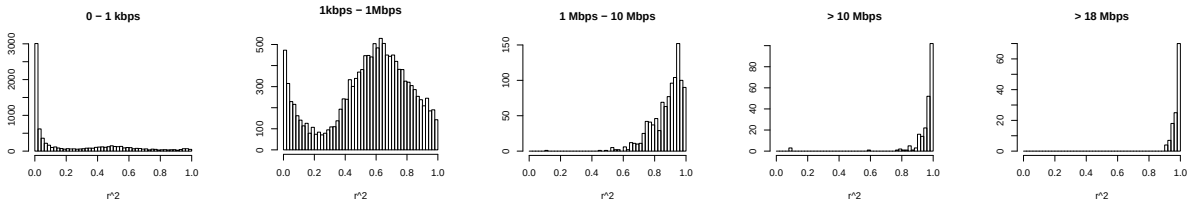

Figure 8: Testing Gaussianity: Distribution of r^2 values for OD pairs of different traffic volumes.

linear. Goodness of fit with respect to this can be calculated by the linear correlation coefficient $r(x, a)$, defined in (1), and the value r^2 is used as a measure of the goodness of fit, an approach used in [10] and in our earlier works [7, 8]. In [9] the authors studied this method and found that although simple, it is sufficiently accurate to determine the validity of the Gaussian assumption. They note that when $r^2 > 0.9$ then also the more complex Kolmogorov-Smirnov test usually supports the assumption that the traffic is Gaussian.

In Figure 7 the size of the OD pair traffic volume (bits per second) is plotted against the goodness of fit value r^2 of the Gaussian assumption. We can see from the figure that the larger flows are always close to Gaussian, with r^2 values easily over 0.90. The largest OD pair with $r^2 < 0.90$ has traffic volume of 17.5 Mbps. The vertical line in the figure is located at 10 Mbps, which

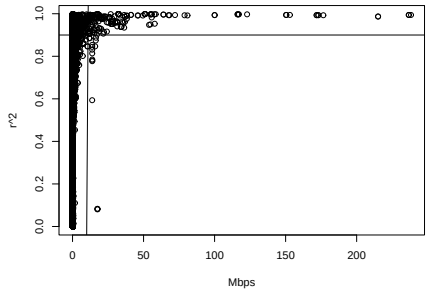

Figure 7: Testing Gaussianity: Goodness of fit values r^2 as a function of OD pair traffic volume.

seems to be an approximate threshold after which an overwhelming majority of the OD pairs have $r^2 > 0.90$, with $r^2 > 0.98$ for the many of the OD pairs, as seen in the histogram of Figure 8. For OD pairs of size 1 Mbps to 10 Mbps there is still a lot of Gaussian traffic, while for OD pairs smaller than 1 Mbps there is not any Gaussian behavior observable. For smallest OD pairs the fit is almost always near zero, as these are typically flows that have one or few bursts of traffic and are idle the rest of the time.

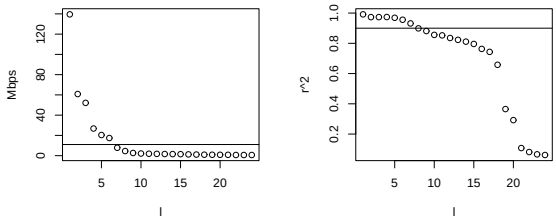

Figure 9: Testing Gaussianity: Average OD pair traffic volumes and goodness of fit values r^2 as a function of prefix length l. Direction d_0.

In Figure 9 the average OD pair traffic volumes and the average r^2 values are shown as a function of prefix length. The average is taken over those largest OD pairs that comprise 80 percent of total traffic. For the link direction d_0, depicted in the figure, the first six cases, with prefix lengths from 1 to 6, have an aggregation level high enough so that their average traffic volume is over 10 Mbps, and the r^2 values for the first seven cases exceed

0.9. For the d_1 direction, the first six are over 10 Mbps and the same six are over 0.9 while the the seventh is almost exactly 0.9. In general, the ten megabit threshold seems to approximately apply also for averages. An average of 10 Mbps implies that the goodness of fit is better than 0.90. However, in both directions the values decline rather slowly from good to reasonable to adequate until a steep drop occurs from the adequate values to the bad values between network prefixes of 15 and 20 bits. While Figure 9 is in linear scale and fails to depict any observable change in the mean flow size in this region, Figure 4, in logarithmic scale, shows a steep drop in the maximum size of the OD pair.

To summarize, while it is impossible to set a concrete threshold, it seems that in our data majority of the OD pairs with at least 10 Mbps of traffic are fairly Gaussian.

6 Mean-variance relation

In traffic matrix estimation the spatial mean-variance relation is used to obtain necessary extra information about an otherwise underdetermined problem. A functional relation is assumed between the mean $\boldsymbol{\lambda}$ and the variance $\boldsymbol{\Sigma}$ of an OD pair's traffic volume.

The spatial mean-variance relation is a key assumption in many traffic matrix estimation techniques [1, 11, 12, 13], but evidence of its validity is contradictory. Cao et al. found it to be sufficiently valid to justify using it, but their study is of a local area network, which is not representative of backbone traffic. Gunnar et al. [6] find the relation valid in study of a Global Crossing backbone, while Soule et al. consider the validity not sufficient in their study [14]. We found the relation to hold moderately in the Funet network, with average goodness

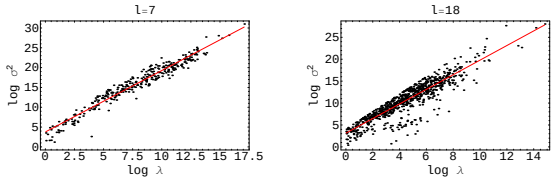

Figure 10: Mean variance relation in log-log scale. Left: $r^2 = 0.95$, right: $r^2 = 0.80$.

of fit value around $r^2 = 0.80$ [8]. That study, however, was done with extremely high resolution, leading to very small traffic volumes. Now we have the measurement data available to extend the results for larger aggregation levels, which are more relevant as typical traffic matrix estimation environment is a backbone network with large traffic volumes.

The commonly used power law relation can be written as

$$\boldsymbol{\Sigma} = \phi \cdot \mathrm{diag}\{\boldsymbol{\lambda}^c\}.$$

The power law relation for the OD pair i is $\sigma_i^2 = \phi \cdot \lambda_i^c$, and its logarithm is $\log \sigma_i^2 = c \log \lambda_i + \log \phi$. Thus, if the relation held, the points would fall on a line with slope c and intercept $\log \phi$ in the log-log scale. This is a simple linear regression model and we can measure the validity of the mean-variance relation with the linear correlation goodness of fit value r^2 used in the previous section.

For each prefix length mean and variance are calculated for each one hour period in the 24 hour trace. In Figure 10 the values are depicted for one selected hour and two selected prefix lengths, with one point in the plot representing the mean and the variance of one OD pair for that hour. For a longer prefix ($l = 18$) $r^2 = 0.80$, which is in line with previous results. It can be seen that the the values defer significantly more from the regression line making the fit worse. However, for shorter prefix ($l = 7$), depicted in the same Figure, the fit is much better, about $r^2 = 0.95$.

In Figure 11 the average goodness of fits values are shown as a function of the network prefix length l. As the prefix length gets longer, there are more OD pairs, with the average size of an OD pair obviously getting smaller. Recall that the average OD pair sizes for different prefixes are shown in Figure 9. For the longer prefixes the fit of the mean-variance relation is around 0.75 to 0.80. As the resolution gets coarser, the goodness of fit values improve to over 0.90, in some cases as high as 0.95. The OD pair traffic volumes at these aggregation levels are still less than 100 Mbps, and as the growth is approximately linear as a function of the aggregation level, we may conclude that for larger traffic flows the fit is at least as good, probably better.

Table 1 shows the values of the exponent parameter c with different aggregation levels. It can be said that the parameter stays relatively constant and that the values

Table 1: Estimates for the mean-variance relations exponent parameter c for different prefix lengths l.

l	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
c	1.64	1.60	1.60	1.60	1.66	1.71	1.72	1.76	1.77	1.73	1.75	1.76	1.73	1.71	1.67	1.66	1.71

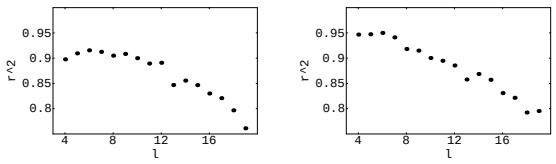

Figure 11: Testing mean-variance relation: Goodness of fit values r^2 as a function of prefix length l. Directions d_0 on the left side, d_1 on the right side.

fall between the results reported for parameter values in other networks [1, 6, 14].

We can conclude that there is a clear dependency of the mean-variance relation fit and the aggregation. Most importantly, there is a strong functional mean-variance relation for the cases where aggregation level is high.

7 Conclusion

In this paper we have analyzed the origin-destination pair traffic in the Funet network, and in particular the effects that spatial aggregation has on these characteristics.

Gaussian assumption holds better when the aggregation level is higher. An approximate threshold, after which all OD pairs are at least fairly Gaussian, would appear to be around traffic volumes of 10 to 20 Mbps. This means that for many traffic engineering and traffic modeling tasks where we consider much larger traffic flows the Gaussian assumption is justified, but it probably cannot be used for cases with smaller traffic volumes due to low aggregation level.

The diurnal variation of the OD pairs follow the diurnal pattern of total traffic more closely when the aggregation level is higher. However, there is not a clear cut boundary as in the Gaussianity assumption, so it is difficult to say anything concrete. We can point out, though, that it would be ill-advised to assume in any scenario that diurnal patterns are similar for all OD pairs, or that busy hours of different flows would coincide.

We validated the spatial power law assumption between mean and variance of the OD pairs. Particularly with large aggregation levels it holds well. This is an essential result concerning traffic matrix estimation techniques which rely on this very assumption. Our results also show that the exponent parameter remained about constant regardless of the aggregation, and was within the range of values obtained for it in literature.

To conclude, we can state that the more aggregated the traffic becomes, the more well behaved it is in general, in the sense that the assumptions studied hold better.

References

[1] J. Cao, D. Davis, S. V. Wiel, and B. Yu, Time-varying network tomography, Journal of the American Statistical Assosiation, Vol. 95, pp. 1063-1075, 2000.

[2] M. Barthelemy, B. Gondran, and E. Guichard, Spatial structure of the internet traffic, Physica A: Statistical Mechanics and its Applications, Volume 319, 2003.

[3] A. Feldmann and A. Greenberg and C. Lund and N. Reingold and J. Rexford and F. True, Deriving Traffic Demands for Operational IP Netowrks: Methodology and Experience, IEEE/ACM Transactions on Networking, 9:3, 2001.

[4] S. Bhattacharyya and C. Diot and J. Jetcheva and Nina Taft, Pop-Level and Access-Link-Level Traffic Dynamics in a Tier-1 POP, Proceedings of ACM Internet Measurement Workshop (IMW), 2001.

[5] A. Lakhina, K. Papagiannaki, M. Crovella, C. Diot, E.D. Kolaczyk, and N. Taft, Structural

analysis of network traffic flows, in SIGMETRICS/Performance 2004, New York, USA, June 2004.

[6] A. Gunnar, M. Johansson, and T. Telkamp, Traffic matrix estimation on a large IP backbone- A comparision on real data, in IMC 2004, Taormina, Italy, October 2004.

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

[8] R. Susitaival, I. Juva, M. Peuhkuri, and S. Aalto, "Characteristics of OD pair traffic in Funet", in ICN 2006, Mauritius, April 2006. (Extended version to appear in Telecommunications Systems)

[9] R. van de Meent, M. Mandjes and A. Pras, "Gaussian Traffic Everywhere?", ICC 2006, Istanbul, Turkey. June 2006.

[10] J. Kilpi and I. Norros, "Testing the Gaussian approximation of aggregate traffic", in 2nd ACM SIGCOMM Internet Measurement Workshop, Marseille, France. 2002.

[11] Y. Vardi, "Network tomography: estimating source-destination traffic intensities from link data", Journal of the American Statistical Association, vol 91, pp 365–377, 1996.

[12] G. Liang and B. Yu, "Pseudo Likelihood Estimation in Network Tomography", IEEE Infocom, 2003.

[13] I. Juva, S. Vaton and J. Virtamo, "Quick Traffic Matrix Estimation Based on Link Count Covariances", in proceedings of ICC 2006, Istanbul, Turkey. 2006.

[14] 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. 2004.