Internet Traffic Mid-Term Forecasting: a Pragmatic Approach using Statistical Analysis Tools

Rachel Babiarz Jean-Sebastien Bedo

France Telecom R&D Division, Innovation Economics Laboratory,
38-40 rue du Général Leclerc, 92794 Issy-les-Moulineaux Cedex 9, France
{rachel.babiarz,jeansebastien.bedo}@rd.francetelecom.com

Abstract. Network planning is usually based on long-term trends and forecasts of Internet traffic. However, between two large updates, telecommunication operators deal with resource allocation in contracts depending on the mid-term evolution of their own traffic. In this paper, we develop a methodology to forecast the fluctuations of Internet traffic in an international IP transit network. We do not work on traffic demands which can not be easily measured in a large network. Instead, we use link counts which are much simpler to obtain. If needed, the origin-destination demands are estimated *a posteriori* through traffic matrix inference techniques. We analyze link counts stemming from France Telecom IP international transit network at the two hours time scale over nineteen weeks and produce forecasts for five weeks (mid-term). Our methodology relies on Principal Component Analysis and time series modeling taking into account the strain of cycles. We show that five components represent 64% of the traffic total variance and that these components are quite stable over time. This stability allows us to develop a method that produce forecasts automatically without any model to fit.

Keywords: IP international transit network, traffic forecasting, principal component analysis, time series modeling, strain modeling.

1 Introduction

Forecasting the end to end traffic profiles of an IP network is very important in order to deal with traffic engineering tasks like new resources planning or network design. Whereas these end to end traffic demands can be directly obtained for traditional telecommunication networks based on circuit switching, it is more difficult with packet-based routing which is used in Internet. Hence, it relies on the IP protocol which does not include an accounting mechanism. The authors of [1] highlight these difficulties. Tools like Netflow from Cisco have been developed to directly measure the traffic demands, but these measurements are quite difficult to obtain in an accurate and exhaustive manner. These tools are based on sample measurements, use a lot of network resources and then cannot be activated on all routers. So, we do not have historical data on the end to

end traffic demands. The only easily available historical data are the amount of traffic exchanged between adjacent routers (link counts) that can be obtained through the Simple Network Management Protocol (SNMP). The link counts correspond to the sum of several users traffic demands entering the network into one edge router and exiting at an other edge router. To obtain the traffic origins and destinations, a lot of traffic matrix inference techniques have been developed these last few years. They use link counts and routing schemes to estimate the end to end demands of the network. [2–5] give a quite complete overview of the recent work done by the research community on this topic.

In this paper, our goal is then to predict the link counts. We need online forecasting to increase our reactivity and flexibility. So we propose a completely automatic technique. We test our approach on France Telecom IP international transit network. We work on several weeks of SNMP data, at a large time scale (two hours granularity) and do forecasts for a few weeks. The main difficulty is that there are usually many links (next to a thousand) in an international transit network and it is not feasible to fit a model on each link to produce its forecasts. We develop a pragmatic approach to deal with this high dimension and assure its full automation for operational purposes. Crovella et al. show in [6] that Principal Component Analysis (PCA) applied to link counts data can drastically reduce the high dimension into a few components. This is due to high correlations existing between link counts evolutions. We see that these few components stemming from PCA exhibit strong time periodicities that do not change very much. We study the cycle change of shape of the components and propose a new approach to build forecasts for this kind of data based on simple statistical tools. Our forecasting technique has the advantage to be fully automated contrary to classical time series models such as SARIMA which are needed to be fitted in several steps. The contribution of our paper to the field is to validate the PCA forecasting methodology proposed by Crovella et al. on new real traces of a large international backbone IP and the introduction of the study of the shape morphing inside the forecasting methodology itself.

The paper is organized as follows. In section 2, we briefly present previous works relative to traffic forecasting. Then, we describe the data on which we have applied our methodology in section 3. The section 4 details the way we can decompose IP traffic observed on a lot of links simultaneously in elementary shapes thanks to PCA. The forecast techniques we propose are developed in the next section. The last section is devoted to the results forecasts descriptions.

2 Related Work

Forecasting traffic was already an issue for the Plain Old Telephone Service (POTS) ([7, 8]). Most of the processing theories have been extensively used to deal with this problem. But there was no statistical multiplexing, so traffic profiles were much more continuous contrary to IP traffic profiles. Internet traffic forecasting techniques (see for example [9–11]) have mainly addressed local area network and small time scales, such as seconds or minutes, that are relevant for dynamic resource allocation and show the local effects of statistical multiplexing.

In our case, we are interested in international transit network and larger time scales which are more appropriate when doing capacity planning and network design. But long range dependencies effects begin to appear very rapidly as the time scale grows ([12]). The first work dealing with large time scale is described in [13]. In this paper, the authors predict a single value for the entire network using linear time series models which is not sufficient for network planning purposes.

The nearest work to ours is exposed in [14]. The evolution of IP backbone traffic at large time scales are modeled and long-term predicted combining wavelet multiresolution analysis and linear time series models.

3 Data

We collected SNMP data from all the routers of the France Telecom IP international transit network from April 3, 2005 to August 13, 2005. These data represent the total amount of traffic exchanged between all the adjacent routers (link counts) by ten minutes time slots. For this nineteen weeks period, about eight hundred links have been observed between next to two hundred routers. These links are either access links or core links, we do not distinguish between this two kind of links in this paper and do forecasts for both type. Among all the links, about two hundred of them are not active during all the period or have a negligible traffic amount. We do not consider these specific links for the forecasts. As we are interested in doing forecasts for network planning purposes, we average our traffic measurements across two hours intervals. We do not average our data on a larger interval because we want to keep the daily periodicities (see the next paragraph) which are an important traffic element useful for the forecasts buildings. Two hours granularity is then a good compromise.

We observe two types of traffic behavior in our link counts data. There are link counts that exhibit strong daily and weekly periodicities reflecting traditional human activities. This behavior corresponds to the largest link counts and represent the majority of the total traffic. Other link counts are bursty, representing occasional spikes or dips of traffic. They mainly correspond to small link counts. We show an example of these two types of traffic behavior in Figure 1. The period of time is intentionally reduced to allow to distinguish the cycles.

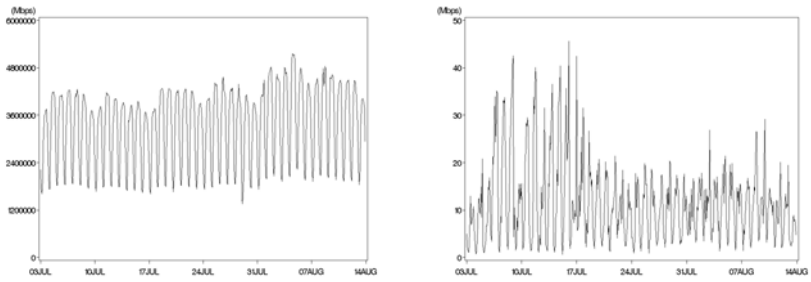

Fig. 1. Examples of a cyclic (left) and a bursty (right) link count

4 Structural Analysis of Link Counts

4.1 Principal Component Analysis Overview

In this part, we intend to introduce the Principal Component Analysis (PCA) framework but we cannot cover all aspects due to lack of space. For more details, see [6]. The method of PCA is useful to analyze a complex set of many correlated statistical variables $X = \left[X^1, ..., X^p\right]$ into new principal independent components ([15]). PCA works on zero-mean data. The principal components correspond to the eigenvectors of the covariance matrix $X^T X$:

$$X^T X\, v_i = \lambda_i\, v_i \quad i = 1, ..., p \tag{1}$$

λ_i is the eigenvalue corresponding to the eigenvector v_i. Since $X^T X$ is symmetric definite positive, its eigenvectors are orthogonal, the eigenvalues are non-negative real and its trace, corresponding to the total variance of X, is equal to the sum of the eigenvalues, so that λ_i represents the variation part of X captured by the i^{th} eigenvector or principal component. By convention, the eigenvectors are unit vectors and the eigenvalues are sorted from large to small. Thus, the first eigenvector v_1 captures the largest variation part of X, the second eigenvector v_2 the second largest variation part, and so on. The projection of X on v_i represents the coordinate or contribution of X on the i^{th} principal component, this vector can be normalized to unit length by dividing by $\sqrt{\lambda_i}$:

$$u_i = \frac{X\, v_i}{\sqrt{\lambda_i}} \quad i = 1, ..., p \tag{2}$$

u_i is then a linear combination of the initial variables, it is usually named a score. By performing PCA, we decompose X into an optimal sum of unit rank matrices (product of a line vector by a column vector):

$$X = \sum_{i=1}^{p} \sqrt{\lambda_i}\, u_i\, v_i^T \tag{3}$$

If we consider only the first q principal components and scores, we obtain the best approximation of X with q elements. This approximation can be computed by taking p equal to q in Equation 3. In the next part, we apply the technique of PCA on our link counts data.

4.2 PCA Application on Link Counts

The matrix X, defined in the previous section, now represents the T measurements of traffic on the nineteen weeks observed period and on the L links (variables) of our network. As we have seen in section 3, the traffic data exhibit strong daily and weekly periodicities in majority. To respect these properties, we decide to perform PCA week by week. We center and reduce the traffic evolutions for each link and per each week in order to consider traffic profile instead of raw

traffic. This allows to give as much importance to all the links in the network. We obtain, for each week, L new vectors or scores, we call them the eigenlinks in reference to Crovella et al. who call them the eigenflows in [16] when PCA is applied on OD flows traffic. The eigenlinks represent the elementary shapes of the link counts. The first eigenlink captures the strongest time variation common to all links, the second eigenlink the next strongest, and so on. We represent in the figure below the variation part (cumulative percentage) captured by the first twenty eigenlinks in the form of a scree plot. A scree plot shows the sorted eigenvalues, from large to small, as a function of the eigenvalue index. Each curve corresponds to a different studied week.

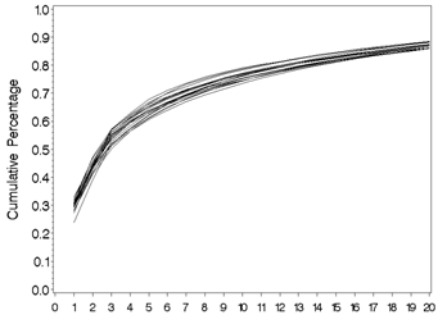

Fig. 2. Scree Plot for Link Counts

We can see that the vast majority of link counts variability is explained by the first few eigenlinks, whatever the week we consider. This is due to the fact that the majority of the link counts show the same strong daily and weekly periodicities. So, the link counts form a structure with effective dimension much lower than the total number of links.

The first nine eigenlinks are represented in Figure 3. Once again the period of time is reduced for readability reasons. The first five eigenlinks exhibit strong daily and weekly periodicities. This is not surprising as these properties concern the majority of link counts data. Their periods are different and their peaks of traffic do not appear at the same time. This is due to the different time zones covered by our network. The next eigenlinks do not reflect a particular shape, they are more bursty. In [16], a taxonomy of eigenflows is realized using heuristics based on their periodogramm and their standard deviations. The authors show that the eigenflows can be separated in three categories in a quantitative way: deterministic, spike and noise. By using the same quantitative heuristics and taxonomy, only the first five eigenlinks are characterized as deterministic. We then decide to only use the first five deterministic eigenlinks to build our forecasts for all the link counts of the France Telecom international transit network. We reduce the dimension from about six hundred to five thanks to PCA. These five elements represent 64%, in mean over the studied weeks, of the link counts total variation.

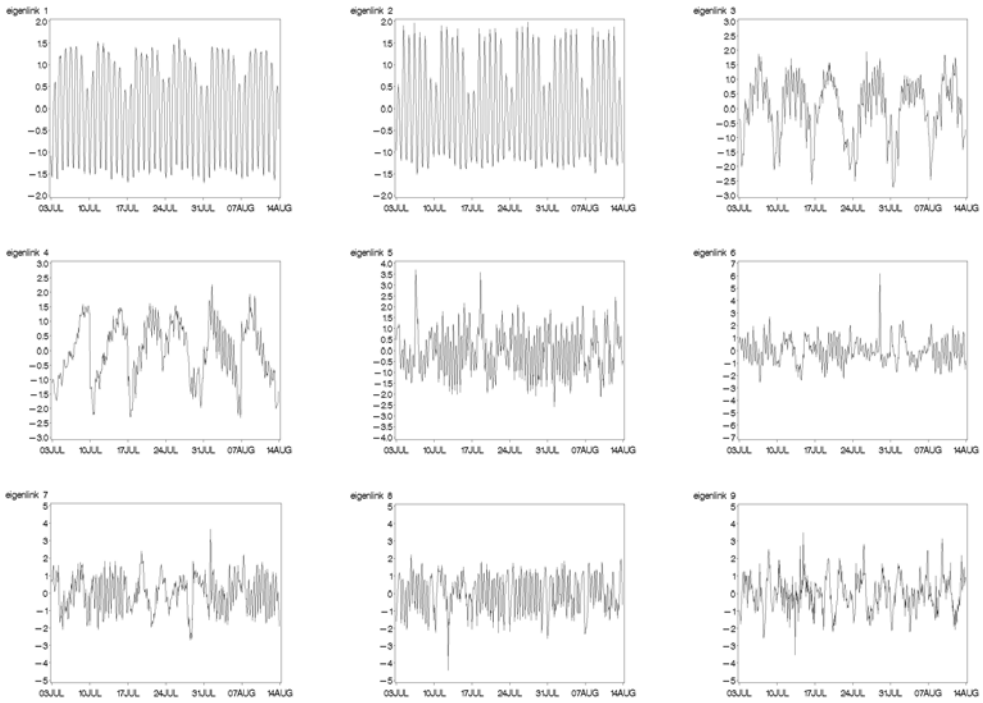

Fig. 3. First Nine Eigenlinks

If we represent the first five eigenlinks by superposing the weeks between them (Figure 4), we can see that the link counts structure is quite stable over time. We use this property to develop our forecasting method in the next section.

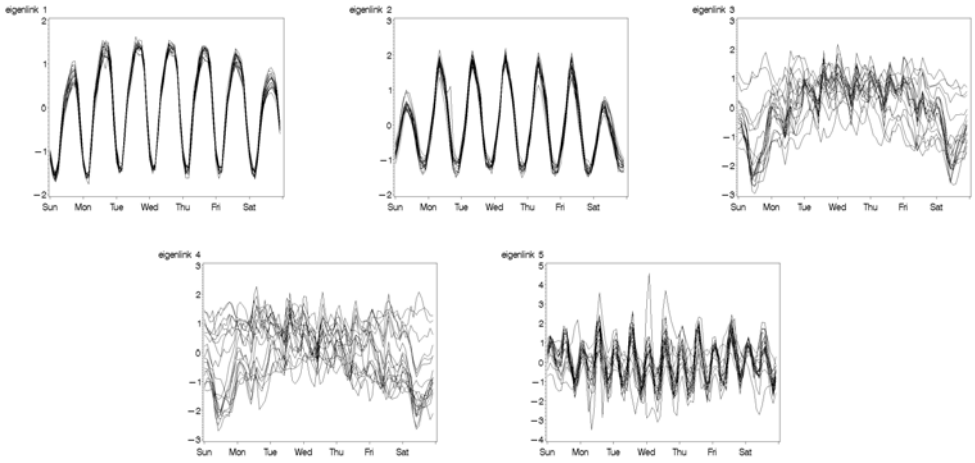

Fig. 4. First Five Eigenlinks superposed week by week

5 Forecasting Techniques

We propose to build forecasts for the deterministic eigenlinks time series and then to project these forecasts on the adequate eigenvectors to obtain all the link counts forecasts via Equation 3 with q equal to the number of deterministic eigenlinks kept. It can be compared to a generalized deseasonalization method ([17]). We discuss the adequate eigenvectors choice further. As we want to develop a methodology that can be fully automated for network planners, we do not use classical linear time series models such as SARIMA models which are needed to be fitted manually one by one (see [18] for details on the Box-Jenkins methodology to fit SARIMA models). We propose a pragmatic approach based on basic statistical tools such as mean and standard deviation and using the time stability of the link counts structure stemming from the first deterministic eigenlinks we have already observed in the previous section.

The study of the ratio between a sequence mean and its standard deviation is a good way to measure the dispersion of this sequence. Indeed, the higher this ratio is, the more stable the serie is. We use this criteria to quantify the stability of each deterministic eigenlink time serie. We then build the following indicator:

$$Y_{t_1}^t = \frac{Mean_{\,i}\,(X_t^i - X_{t_1}^i)}{Standard\,Deviation_{\,i}\,(X_t^i - X_{t_1}^i)}\,,\,\forall\, t_1 \neq t \tag{4}$$

where t represents the sub-measurements of a cycle. In our case, the cycle corresponds to a week, so t varies from Sunday 0h-2h to Saturday 22h-0h, i.e. 84 sub-measurements in all. X_t^i represents the traffic volume for the sub-measurement t for the week i. If $Y_{t_1}^{t_2}$ is high, it means that the cycle do not change its shape between the sub-measurements t_1 and t_2. Then, if we want to forecast the value of X_{t_2} for the next cycle $K+1$ from the K observed cycles, we can do as follows:

$$X_{t_2}^{K+1} = X_{t_1}^K + Mean_{\,i=1}^{\,K}\,(X_{t_2}^i - X_{t_1}^i) \tag{5}$$

We generalized Equation 5 by considering all the sub-measurements t of a cycle:

$$X_t^{K+1} = \sum_{t_1 \neq t} weight_{\,t_1}^{\,t}\,[\,X_{t_1}^K \,+\, Mean_{\,i=1}^{\,K}\,(X_t^i - X_{t_1}^i)\,],\,\forall\, t \tag{6}$$

where

$$weight_{\,t_1}^{\,t} = \frac{(Y_{t_1}^t)^2}{\sum_{t_1 \neq t}(Y_{t_1}^t)^2} \tag{7}$$

Our prediction formula defined by Equation 6 consists of computing the prediction for a sub-measurement t of a cycle by a weighted sum of all the other sub-measurements. The introduction of a weight stemming from the cycle stability indicator defined by Equation 4 allows to consider more importance to the closest sub-measurements to the sub-measurement for which we compute the

forecast in terms of strain between cycles. We use this technique to compute the deterministic eigenlinks forecasts. The results are given in the next section.

Once the eigenlinks forecasts are obtained, the final forecasts for all the link counts are computed by projecting these forecasts on the eigenvectors. As, we have performed PCA on our data week by week, we have a set of eigenvectors and the corresponding eigenvalues for each week or cycle. We propose to use the eigenvectors and eigenvalues stemming from the last observed cycle as we have seen that the structure is stable over time.

6 Forecasting Results

6.1 Deterministic Eigenlinks Forecasts

We divide our observation period in two parts: an estimation period from April 3, 2005 to July 9, 2005 (14 weeks) on which we develop our forecasting techniques and an evaluation period from July 10, 2005 to August 13, 2005 (5 weeks) from which we compare our forecasts results to the real data. Figure 5 shows the forecasting results we obtain for the first five eigenlinks using our methodology (Equation 6). The forecasts are in dashed lines and the real values in plain lines.

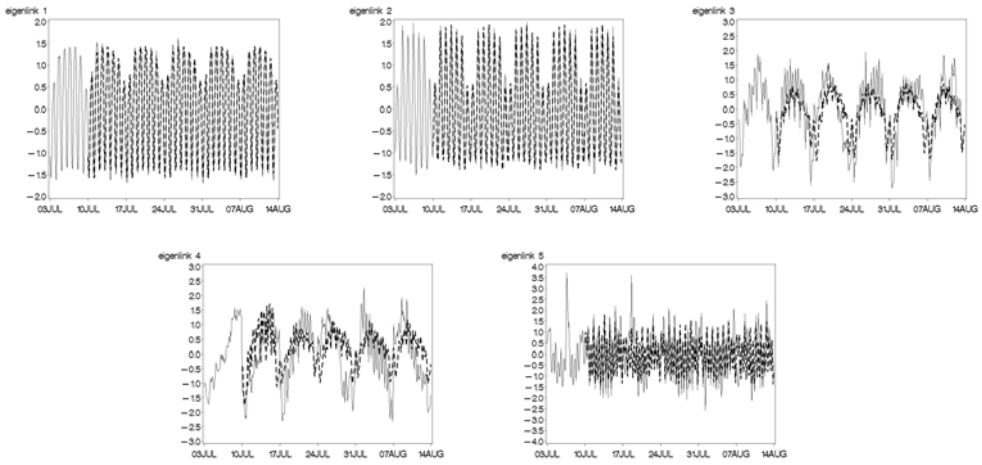

Fig. 5. First Five Eigenlinks Forecasts

We can notice from these graphs that the first components are quite well forecasted even if their shape is smoothed by the technique. However the forecasts for the fourth component are distorted compared to real data. This is mainly due to the high strain between sequential cycles on this component.

We compare these forecasting results with those obtained when SARIMA models are fitted. The following models have been fitted: $SARIMA\,(1,0,1)_{1,12,84}$ for the first four eigenlinks and $SARIMA\,(1,0,1)_{1,6,12,84}$ for the fifth eigenlink. For these comparisons, we compute the median of the absolute relative error

(Equation 8) for each eigenlink. X corresponds to the real values and X' to the forecasts. We do not compute the mean because of some extreme values due to abnormal traffic behavior. We can see in Table 1 that our technique gives in general slightly better results than SARIMA models. We do not pretend to give better results than any SARIMA model. The SARIMA models we fit for the comparisons are choosen with the same orders for convenience (we need an automated method). In addition, the method we proposed does not rely on iterative algorithms (contrary to SARIMA models) leading to a lower computational burden.

$$Median\left(\frac{|X' - X|}{|X|}\right) \tag{8}$$

	Our Method	SARIMA
Eigenlink 1	0.1108	0.1046
Eigenlink 2	0.1324	0.1370
Eigenlink 3	0.6408	0.6787
Eigenlink 4	0.8144	0.9258
Eigenlink 5	0.5990	0.6355

Table 1. Comparison Results (Median Error) between our method and SARIMA models

6.2 Link Counts Forecasts

We remind that the forecasts for link counts are obtained thanks to Equation 3 with q=5:

$$X' = \sum_{i=1}^{5} \sqrt{\lambda_i}\, u_i\, v_i^T \tag{9}$$

λ_i and v_i correspond respectively to the i$^{\text{th}}$ eigenvalue and the i$^{\text{th}}$ eigenvector stemming from the PCA application on the last week of the estimation period, and u_i is the i$^{\text{th}}$ forecasted eigenlink over the evaluation period. Then, X' is a matrix containing the forecasts for all the link counts. As we have centered and reduced the data to zero mean and unit standard deviation before applying PCA, the final forecasts results have to be rescaled by the mean and the standard deviation of each link count respectively. We use the mean and standard deviation of link counts observed on the last week of the estimation period.

Besides the median absolute relative error defined in Equation 8, we also compute the relative errors over the median (Equation 10) and over the maximum (Equation 11), the traffic median or maximum being indicators usually used for network design.

$$\frac{Median(X') - Median(X)}{Median(X)} \tag{10}$$

$$\frac{Max(X') - Max(X)}{Max(X)} \tag{11}$$

where X corresponds to the real values and X' to the forecasts. We compute the three types of error for each link count and each week of the evaluation period. We represent in Figure 6 these errors for all the link counts classified from large to small and for the first (dashed line) and last (dotted line) week of the evaluation period.

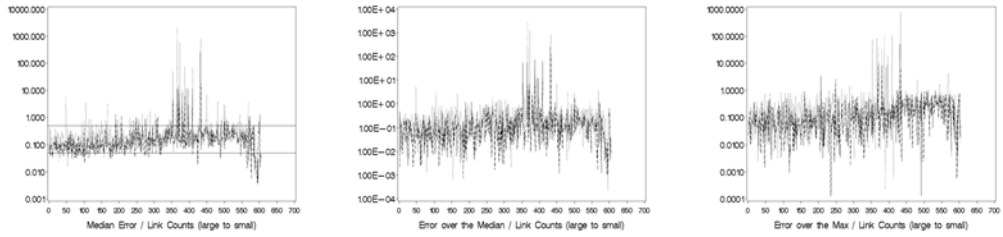

Fig. 6. Forecast Errors

We can see that the majority of link counts have a median error lying between 5% and 50% (horizontal lines). The errors have the same order of height whatever the traffic volume of the link count and are quite stable between the first and last weeks of the evaluation period. The errors peaks mainly correspond to:

- bursty link counts without cycles
- links with a traffic mean which varies a lot between weeks
- anomalies of traffic

We would not have significatively better results with other methods. This is due to the traffic sporadicity which makes this specific error peaks unpredictable. We show examples of these types of behavior in Figure 7. The forecasts are in dashed lines and the real values in plain lines.

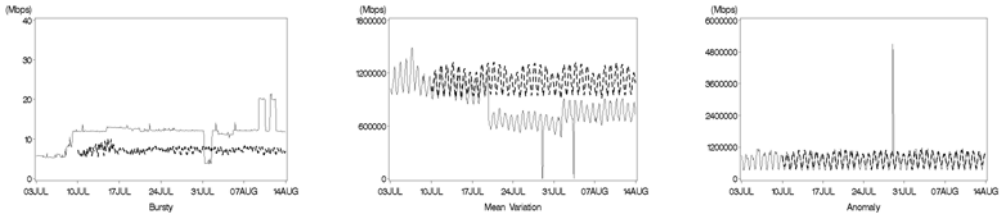

Fig. 7. Examples of Link Counts with bad Forecasts

6.3 Discussion about traffic matrix estimation

In this paper, we have proposed to forecast link counts and then to obtain the end to end traffic demands thanks to traffic matrix estimation techniques based on link counts only. One could wonder how would behave our forecasting method

on historical origin-destination (OD) counts directly. Some internal studies on partial traces have shown that the number of deterministic principal components are slightly the same for link counts or OD counts. In addition, the first components are very similar between links and OD pairs. As a result, traffic matrices forecasts would probably be more precise if we used our forecasting method directly on OD counts since we would not add the approximations of the traffic matrix estimation techniques. Unfortunately, measuring complete OD traffic matrices is often impossible in the case of large IP backbones. And storing historical data without interruptions during weeks concerning traffic matrices is even harder. As a consequence, it is not viable today to rely on forecastings based on direct OD counts for large networks.

7 Conclusion

In this paper, we have developed a pragmatic methodology to predict Internet traffic on all the links of an international IP transit network. Our aim is to obtain these prediction results in a fully automated way in order to be directly operational for network planners who have to deal with several traffic engineering tasks like new resources planning or network design. The low computational burden of the method even allows very reactive on demand forecasts with flexible parameters (period of interest, time scale...). Our method is intentionally simple, based on non-advanced scientific tools, again for automated reasons. In summary, this methodology involves the following steps:

- (1) Principal Component Analysis on the link counts
- (2) Determine the deterministic scores or eigenlinks which are relevant
- (3) Prediction of the deterministic eigenlinks profiles using the study of the cycle change of shape
- (4) Projection of the predicted deterministic eigenlinks on the link counts structure obtained from PCA applied in step (1)

We apply this methodology on data stemming from the France Telecom international transit network over a period of nineteen weeks, between April 3, 2005 and August 13, 2005. Five deterministic eigenlinks out of 600 are kept for summing up the link counts structure. It means a reduction of computational burden by 120. Furthermore, our strain analysis framework is faster than SARIMA techniques. The majority of link counts have a median absolute relative error lying between 5% and 50%. The largest prediction errors concern link counts with a non-constant behavior.

Therefore, through this paper, we validate on real traces from a large backbone IP network the use of the PCA technique to forecast Internet traffic profiles in a large number. We have also developed a new simple method for forecasting periodic traffic profiles based on the study of the variations of cycle shapes between successive periods. It gives us some insights to better understand the underlying trends of IP traffic. Our method can be applied on a longer period by considering the mean trend (instead of the mean of the last cycle) of the

link counts data. The time stability of the link counts structure have also to be controlled, PCA which gives the link counts structure must then be reapplied from time to time.

Acknowledgements

The authors would like to thank Anne-Gaëlle Corrion, Benjamin Petiau, Jean-Luc Lutton and Vincent Martin for their helpful comments and advices on this work.

References

1. K. Papagiannaki, N. Taft, A. Lakhina. A Distributed Approach to Measure IP Traffic Matrices. In ACM IMC, October 2004.
2. A. Medina, N. Taft, K. Salamatian, S. Bhattacharyya, C. Diot. Traffic Matrix Estimation: Existing Techniques and New Directions. In ACM SIGCOM, August 2002.
3. S. Vaton, J.S. Bedo, A. Gravey. Advanced Methods for the estimation of the Origin-Destination Traffic Matrix. In Revue du 25$^{\text{ème}}$ anniversaire du GERAD, 2005.
4. A. Soule, A. Lakhina, N. Taft, K. Papagiannaki, K. Salamatian, A. Nucci, M. Crovella, C. Diot. Traffic Matrices: Balancing Measurements, Inference and Modeling. In ACM SIGMETRICS, June 2005.
5. Y. Zhang, M. Roughan, N. Duffield, A. Greenberg. Fast Accurate Computation of Large-scale IP Traffic Matrices from Link Loads. In ACM SIGMETRICS, June 2003.
6. A. Lakhina, M. Crovella, C. Diot. Diagnosing Network-Wide Traffic Anomalies. In ACM SIGCOM, August 2004.
7. A. Passeron, E. Etve. Modelling Seasonal Variations of Telephone Traffic. In ITC, 1983.
8. P. Chemouil, B. Garnier. An Adaptive Short-Term Traffic Forecasting Procedure using Kalman Filtering. In ITC, 1985.
9. S. Basu, A. Mukherjee, S. Klivansky. Time Series Models for Internet Traffic. In IEEE INFOCOM, March 1996.
10. C. You, K. Chandra. Time Series Models for Internet Data Traffic. In IEEE LCN, October 1999.
11. A. Sang, S. Li. A Predictability Analysis of Network Traffic. In IEEE INFOCOM, March 2000.
12. W. Leland, M. Taqqu, W. Willinger, D. Wilson. On the Self-Similar Nature of Ethernet Traffic (Extended Version). In IEEE ACM Transactions on Networking, Vol. 2, No. 1, pp. 1-15, February 1994.
13. N. K. Groschwitz, G. C. Polyzos. A Time Series Model of Long-Term NSFNET Backbone Traffic. In ICC, May 1994.
14. K. Papagiannaki, N. Taft, Z. Zhang, C. Diot. Long-Term Forecating of Internet Backbone Traffic: Observations and Initial Models. In IEEE INFOCOM, April 2003.
15. H. Hotelling. Analysis of a complex of statistical variables into principal components. In Journal of Educational Psychology, Vol. 24, pp. 417-441, 1933.
16. A. Lakhina, K. Papagiannaki, M. Crovella, C. Diot, E.D. Kolaczyk, N.Taft. Structural Analysis of Network Traffic Flows. In ACM SIGMETRICS, June 2004.
17. S. Hylleberg. Seasonality in Regression. Academic Press, Orlando, FL.
18. P. Brockwell, R. Davis. Introduction to Time Series and Forecasting. Springer, 1996.