A Time Series Model of Long-Term NSFNET Backbone Traffic

Nancy K. Groschwitz and *George C. Polyzos*

{groschwi,polyzos}@cs.ucsd.edu

Computer Systems Laboratory
Department of Computer Science and Engineering
University of California, San Diego
La Jolla, CA 92093-0114

Abstract

We used time series analysis to create detailed forecasts of future NSFNET backbone traffic. The resulting ARIMA model made quite accurate forecasts of traffic levels up to a year in advance. It appears that the model can make reasonable predictions for two or more years into the future, suggesting that ARIMA modeling has great promise as a tool for long-range NSFNET forecasting and planning.

1 Introduction

In planning for the future needs of any complex system, accurate forecasting of the workload is important to assess future capacity requirements, and to plan for changes. The research reported here investigates a method for creating accurate, detailed forecasts of future NSFNET backbone traffic.

The NSFNET is the cross-country backbone of the Internet. The Internet is a three-level hierarchy consisting of the NSFNET backbone, a set of regional (or mid-level) networks that connect to NSFNET sites, and thousands of campus or access networks that connect to the regional networks. In the years examined by this research (1988 through 1993), the NSFNET backbone gradually evolved from a T1-speed (1.544 Mbps) to a T3-speed (44.736 Mbps) network. Each backbone node is connected to several other nodes, typically 2–4. The nodes are responsible for packet switching, routing, and data collection [1].

The most dominant feature of the backbone traffic over the last few years is the overall increase in volume. While a rough estimate of future traffic might be obtained by simply fitting a smooth curve to the data, this method ignores a great deal of information. For example, the traffic may have seasonal components that greatly affect the traffic levels at any given time, but that are obscured by the curve-fitting method. A more detailed model would take these seasonal trends into account.

One of the advantages of such a model would be that, due to its better match to the data and more precise predictions, deviations from previous trends could be spotted more quickly, and new and revised predictions could be generated. A detailed forecast of yearly traffic patterns also allows for more accurate planning and better decisions. If hardware changes are required at some point in a coming year, the model could estimate when traffic levels are likely to be lowest. If a particular backbone link is approaching maximum capacity, a detailed model of the traffic patterns on that link could predict when the capacity is likely to be exceeded. If the traffic on a less busy link is growing more rapidly than on another, busier link, models of each could predict when their usage levels would cross (which might have an impact on routing decisions).

Finally, a more accurate model may allow for reasonable predictions several years into the future. Given the enormous growth rate of traffic on the backbone, the ability to make forecasts two or more years in advance has great advantages for planning for future requirements.

2 Overview of Time Series Models

The first requirement for an adequate model of NSFNET backbone traffic is that it must be stochastic, not deterministic. There are many factors affecting the amount of traffic on the NSFNET, most of which cannot be measured or identified. To predict probable future traffic, the best available basis is an analysis of previously observed traffic patterns. Because we want to examine changes in the traffic over time, the second requirement is that the model must be a *time series* model. Furthermore, the NSFNET traffic data is non-stationary, so the model must be of a form that can accept non-stationary data. A time series model that fits these criteria is the autoregressive integrated moving average process (ARIMA) [2].

The ARIMA model is an extension of a set of time series models called autoregressive (AR), moving average (MA), and autoregressive moving average (ARMA) models. An autoregressive model of order p (denoted by AR(p)), predicts the current value of a time series based on the weighted sum of p previous values of the process plus a random shock. (A *shock* is a random drawing from a white noise process with zero mean and finite variance.) A moving average model of order q (denoted by MA(q)), predicts the current value based on a random shock a and weighted values of q previous a's. If these two models are combined, the ARMA model of order (p, q) predicts the current value of the time series based on p previous values and q previous shocks. The advantage of the ARMA model is that many stationary time series can be modeled with p and q values of 0, 1, or 2.

The AR, MA, and ARMA models all require that the data be stationary. It is not uncommon, however, for a time series to show growth or time-dependent variations that violate this stationarity assumption; the ARIMA model [2] was specifically developed for such non-stationary patterns. A non-stationary series can often be

This research is supported in part by a grant of the National Science Foundation (NCR-9119473).

transformed into a stationary series by differencing the data one or more times. An ARIMA model of order (p,d,q) is simply an ARMA (p,q) model that is differenced d times. An ARIMA(p,d,q) model of time series z_t has the form

$$z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + ... + \phi_p z_{t-p-d}$$
$$+ a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - ... - \theta_q a_{t-q}$$

where the ϕ's are the weights for the AR parameters and the θ's are the weights for the MA parameters.

Simple differencing is sufficient to deal with many kinds of non-stationarity. There may, however, be seasonal cycles that overlay other basic trends in the data, and that are not easily handled by simple differencing. For example, for monthly data over a period of years, the data from a particular month would have a relationship both to the months immediately preceding it, and to the same month in preceding years. To accommodate such effects, a seasonal form of the ARIMA model can be used. It is written as ARIMA $(p,d,q)\times(P,D,Q)_s$, where s is the period of the seasonal pattern. With this model, the original series of length n is first differenced by the length of the period, resulting in a new series of length $n-s$. This series is analyzed according to order (P,D,Q), while the original series is analyzed with (p,d,q); in other words, the model (P,D,Q) links the z's that are s units apart, and the model (p,d,q) links the entire series of z's.

2.1 Fitting an ARIMA Model to the Data

When fitting an ARIMA model to time series data, there are three basic steps, which are used iteratively until a successful model is achieved:

1. *Model identification:* This is the determination of the likely values of p, d, and q for this set of data. Often there will be several plausible models to be examined.

2. *Parameter estimation:* Once a set of possible models has been selected, parameter values are determined for each.

3. *Diagnostic checking:* This involves both checking how well the fitted model conforms to the data, and the use of diagnostic tests that are designed to suggest how the model should be changed, in case of a lack of good fit. The diagnostic tests available for ARIMA checking include examination of: the standardized residuals, the autocorrelation of the residuals, and Box and Jenkin's "portmanteau goodness of fit" statistic. (Chapter 8 of Box and Jenkins [2] contains a complete description of these diagnostics.) Based on the outcome of the diagnostic tests, p, d, or q may be changed, and steps 2 and 3 are repeated.

Once a good fitting ARIMA model has been found by this method, it can be used to make forecasts of the future behavior of the system.

3 Data

The data used in this experiment consists of daily packet totals between all NSFNET backbone nodes, between August 1, 1988 and June 30, 1993. From 1988 through 1990, all data is from the T1 network; from January 1991 to November 1992, the data is the sum of traffic on both the T1 and T3 networks; the T1 network was shut down in November, 1992, and from December 1992 through June 1993 data is from the T3 network only. All data was collected by Merit Network, Inc. as part of its operation and management of the NSFNET backbone. This data is published monthly and is available via anonymous ftp from `nis.nsf.net`.

The traffic data is collected using the Simple Network Management Protocol (SNMP) [3]. The packet counts for each backbone node consist of all packets arriving at the node from its regional networks. Packets arriving from other backbone nodes are not counted.

3.1 Missing Data

All data is missing for the months of July and August, 1989. The T3 data (only) is missing for the months of July, August, and October 1991; T3 at that point was less than 10% of the volume of the T1. The T3 data is also missing for the month of May, 1992, when the T3 and T1 were approximately equal in volume.

There are two options for dealing with missing data: ignore them and omit those data points, or estimate the missing points. Because the ARIMA model examines the pattern of data over time, estimating the missing points and preserving the overall pattern was judged to be more desirable. Missing data values were interpolated by averaging preceding and following data. Because the data showed very strong weekly patterns, the interpolation was always based on a value from the same day of the week as the missing point.

In the few instances where a single day of data was missing from the published monthly reports, interpolation was performed using the data from the appropriate day of the week, seven days before and seven days after the missing day.

3.2 Data Analysis

To simplify data analysis, the daily totals were collapsed into weeks. Because the data began on Monday, August 1, 1988, weeks were defined as beginning on Monday and ending on Sunday. The data ends on Wednesday, June 30, 1993; the final partial week (Monday through Wednesday) was ignored. Weekly totals were used because the creation of a daily ARIMA model (with a period of 365) was computationally very slow (about 24 hours on a DECstation 5000/240), which severely limited the number of model variants that could be examined. (Using weekly data and a period of 52 reduced model creation time to about 5 minutes on the same DECstation.)

All data analysis was done with the S-PLUS$^{(R)}$ statistical package [4].

4 Model Identification

The purpose of the model identification step is to determine likely values for p, d, and q. Once one or more promising sets of (p,d,q) have been identified, the model parameters for those orders can be estimated, diagnostic tests can be run, and the resulting model forecasts can be examined.

The order in which p, d, and q are determined is fixed: first d, the level of differencing, then p, the autoregression, and finally q, the moving average. For each, the general restriction is that the value should be 0, 1, or 2.

Differencing is required to make the data stationary. Figure 1 shows the data before differencing, after differencing once, and after differencing twice. Although most of the non-stationarity is removed by the first differencing, there is an additional effect when differencing is done twice, so, tentatively, we set d to 2. (Later diagnostic tests revealed this to be the correct choice, the model produces a better fit when d is 2.)

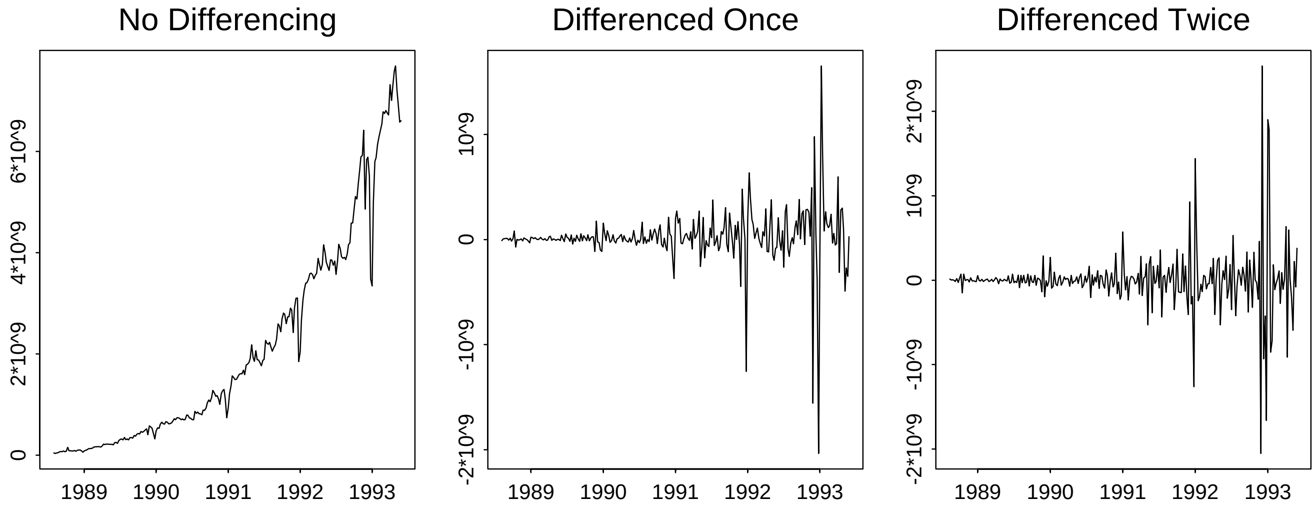

Figure 1: Weekly Packet Totals with and without Differencing

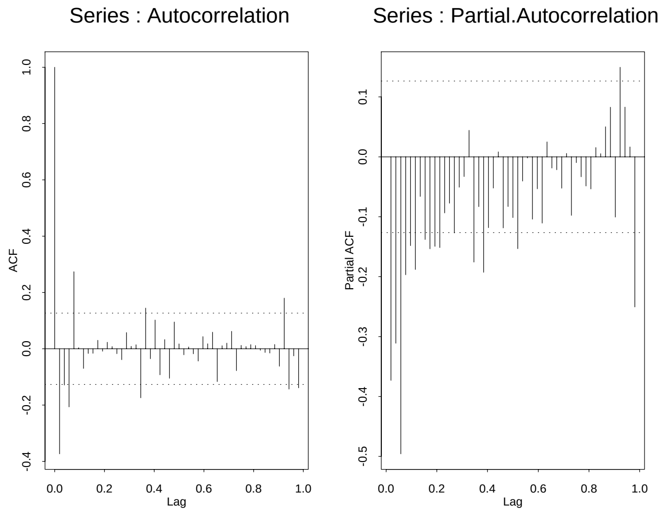

Figure 2: Autocorrelation and Partial Autocorrelation on Differenced Data

Autoregression is determined by examining the pattern of autocorrelations after differencing. If all values except the first are 0, p should be 0. Continually decreasing values indicate a p of 1, and a mixture of increasing and decreasing values mean a p of 2. The left half of Figure 2 shows the autocorrelations; p for these data is clearly 2.

The moving average is determined in a similar fashion, by examining the pattern of the partial autocorrelations, again after the data has been differenced. The meaning of the patterns is the same as for the autoregression. As seen in the right half of Figure 2, the pattern is somewhat ambiguous, possibly indicating a 1, or possibly a 2.

Thus, after the first round of model identification, the tentative choices were (2,2,1) and (2,2,2).

4.1 Diagnostic Checking

The ARIMA diagnostic tests were run on the most promising of the model orders. The three diagnostic tests used are described in Chapter 8 of [2].

Because the data contains evidence of a yearly cycle, diagnostics were also examined for seasonal versions of the promising model orders. Of all model orders tested, the goodness of fit for the seasonal model $(2,2,1)\times(2,2,0)_{52}$ was determined to be the best.

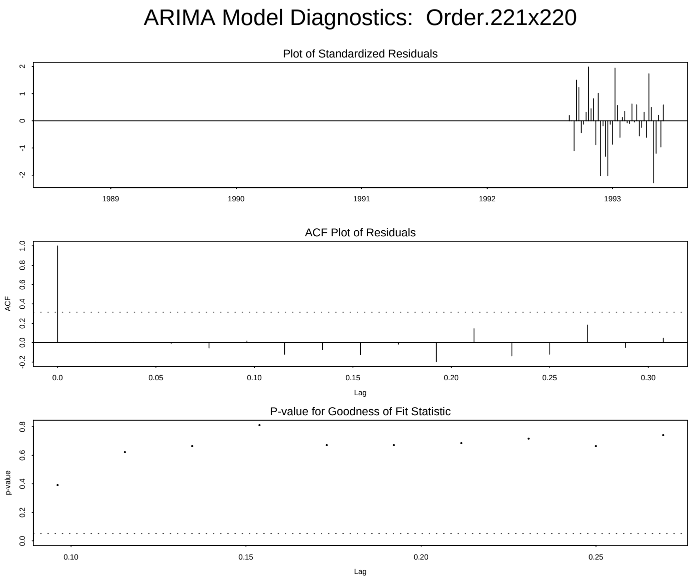

Figure 3: Diagnostic Tests on Model Order $(2,2,1)\text{x}(2,2,0)_{52}$

Figure 3 shows diagnostics results for model order $(2,2,1)\times(2,2,0)_{52}$. The upper plot in Figure 3 shows the standardized residuals. For a good model, the residuals should approximate a Gaussian white noise process, with zero mean, variance of one, and no cycles or patterns; this appears to be a good fit. The middle figure shows an autocorrelation plot of the residuals. The level at which points are significantly different from zero is marked by the dotted line; for a good fitting model, only the first autocorrelation (which is 1.0 by definition) will be significant, which is what is observed here. The lower figure shows the results of the Box and Jenkins portmanteau goodness of fit statistic, a measure of the overall fit of the autocorrelation residuals. A desirable result is that all points should be significantly above zero. The level of statistical significance is marked by the dotted line; all data points are well above the line.

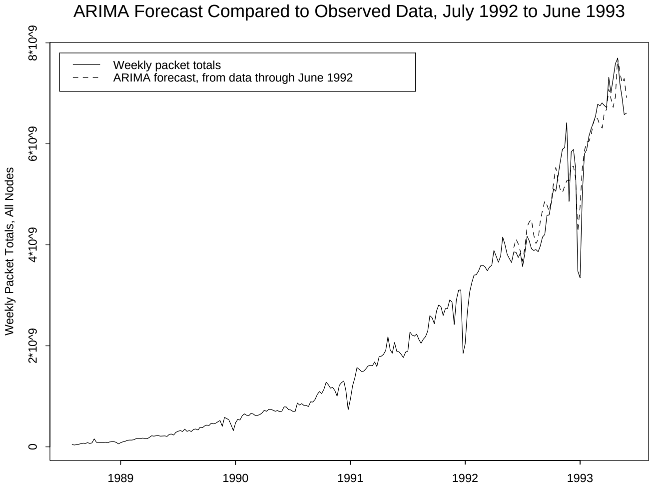

Figure 4: Forecasting with Data from 1988 through June 1992

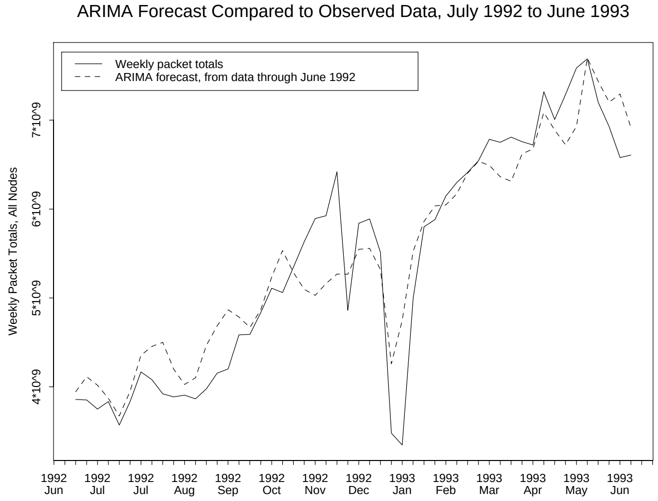

Figure 5: Forecasting with Data from 1988 through June 1992

4.2 The Model Parameters

The estimated parameters (weights) for the model of order (2,2,1)x$(2,2,0)_{52}$ are:

$$\begin{array}{lrl} \text{For (2,2,1)} & \phi_1 = & -0.1768216344 \\ & \phi_2 = & -0.0006852888 \\ & \theta_1 = & 0.9938941 \\ & & \\ \text{For (2,2,0)} & \phi_1 = & -0.2735108 \\ & \phi_2 = & 0.6535305 \end{array}$$

5 Results

The available data is from August 1988 through June 1993. To test the ARIMA model's ability to predict across a full year, the model was given only the data from 1988 through June 1992, and was asked to forecast the subsequent year (i.e., July 1992 through June 1993). This prediction was then compared to the actual data for those 12 months. Figure 4 and 5 show this comparison. Figure 4 shows all data, from 1988 through 1993; Figure 5 shows only the 12 months of the prediction. As can be seen from the figures, the model's prediction was quite accurate over the entire year, although it was too conservative is predicting the size of the oscillations towards the end of 1992.

The model was then used to forecast NSFNET backbone traffic levels for the coming year. All available data, from August 1988 through June 1993, was used for the prediction. Figure 6 shows the forecast for July of 1993 through June of 1994.

A model that could make reasonable forecasts two or more years into the future would have enormous usefulness as a long-range planning tool. In order to evaluate this model's ability to make long-range predictions, the data from 1988 through June of 1992 was used to make a forecast two years into the future (i.e., from July of 1992 through July of 1994). Figure 7 shows this forecast overlaid with the data from Figure 6, which shows a one year forecast based on all available data (1988-1993).

The model's ability to accurately predict one year into the future has been shown by the forecast illustrated in Figures 4 and 5. Thus we can have some confidence in the one-year forecast through July 1994. The fact that the model makes a very similar prediction, based only on data ending two years earlier, suggests that it has real potential as a tool for long-range planning.

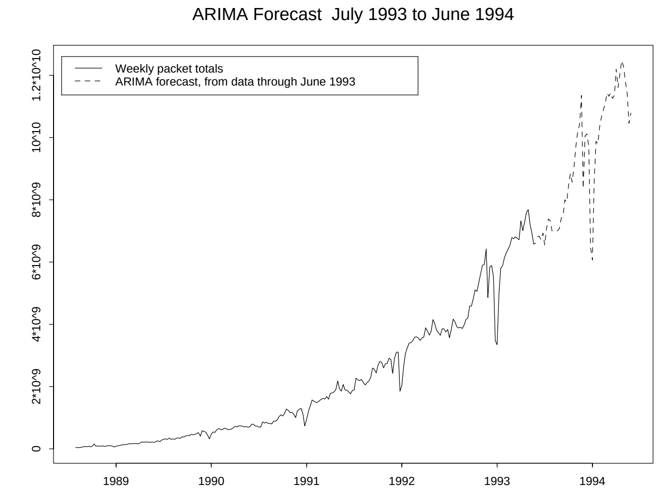

Figure 6: Forecasting with Data from 1988 through June 1993

6 Conclusions

The purpose of this study was to examine the feasibility of using time series analysis to make detailed long-range predictions about NSFNET backbone traffic. In this, it was quite successful. The close match between predicted and observed traffic levels suggests that this approach can be used for long-range forecasts and planning with some confidence.

ARIMA models could also be used to make predictions about the traffic levels on individual NSFNET backbone nodes, rather than on the aggregate. Preliminary model building, however, suggests that it would be necessary to use a separate model order for each node. This should not be surprising, since traffic patterns and growth on individual backbone nodes may be strongly affected by local factors not shared by other nodes, for example, the state of the economy in the region where the node is located, or the addition of a new backbone node that handles traffic formerly handled by this node.

However these examples of outside factors do point to one ultimate limitation of the ARIMA model approach,

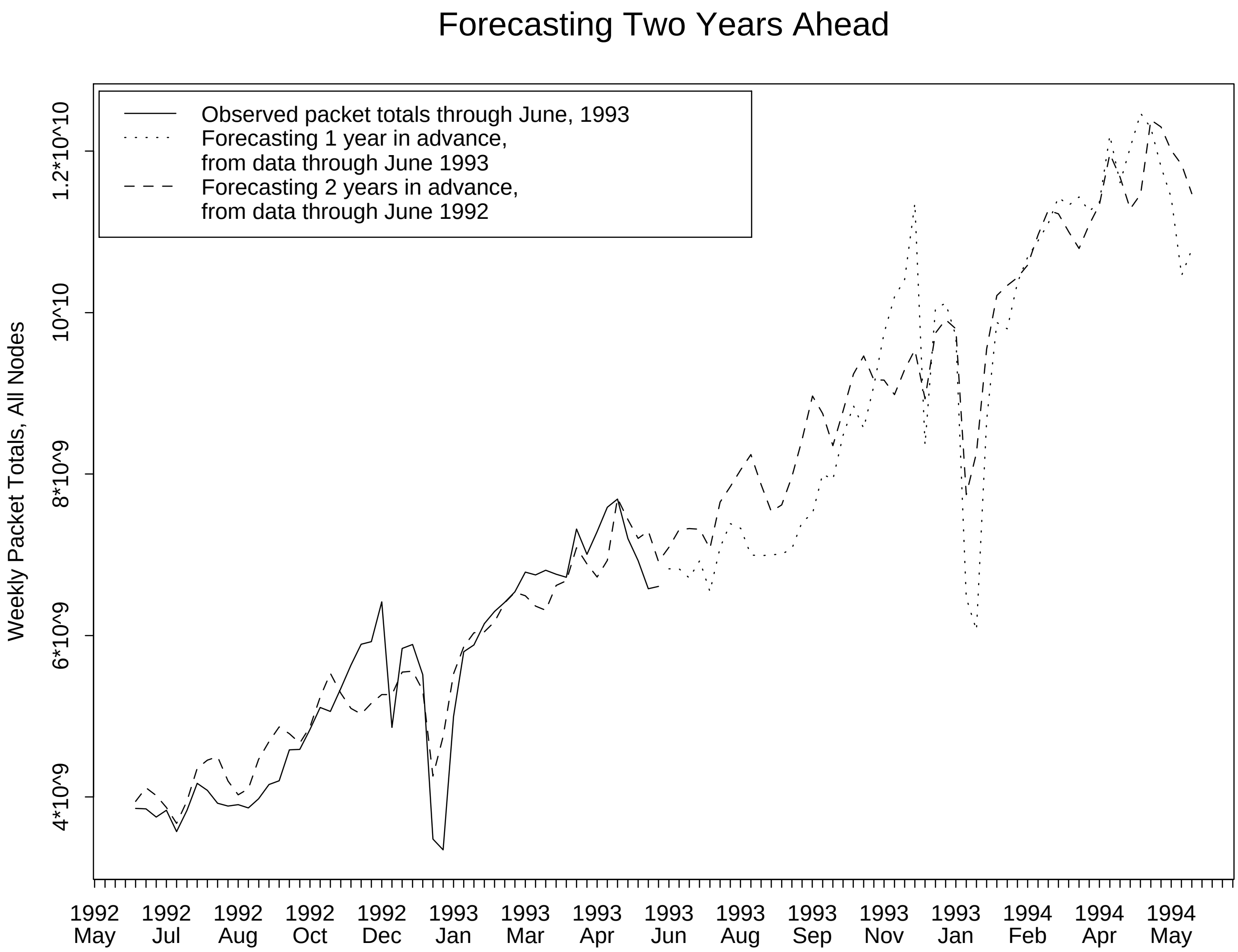

Figure 7: Comparison of One Year and Two Year Forecasts

which is that all predictions are necessarily made on the sole basis of previous data. Such models cannot, by their very nature, take any account of outside forces that may fundamentally change the pattern of the data. As access to the Internet becomes more widespread, a greater proportion of users may be from commercial, rather than academic, institutions; the pattern of use over a calendar year is likely to be quite different for a business than for a university. As new multimedia applications come into wider use, the extremely high volume generated by such applications may affect traffic patterns. Such factors as new technologies, new government regulation, or changes in the national economy may have significant effects that cannot be predicted by this approach.

Acknowledgments

The authors wish to thank Hans-Werner Braun and Kim Claffy for their assistance.

References

[1] K. Claffy, G. C. Polyzos, and H.-W. Braun, "Traffic Characteristics of the T1 NSFNET Backbone," Proceedings *IEEE INFOCOM'93*, pp. 885-892, March 28 - April 1, 1993.

[2] G. Box and G. Jenkins, *Time Series Analysis: Forecasting and Control*, Holden-Day, San Francisco, CA, 1970.

[3] J. Case, M. Fedor, M. Schoffstall, and C. Davin, "Simple Network Management Protocol (SNMP)," Internet Request for Comments Series, RFC 1157, 1987.

[4] *S-PLUS User's Manual*, Statistical Sciences, Inc., Seattle, WA, 1991.