



NORTHWESTERN UNIVERSITY

Computer Science Department

**Technical Report
NWU-CS-02-11
January 12, 2003**

Network Traffic Analysis, Classification, and Prediction

Yi Qiao

Peter Dinda

Abstract

This paper describes a detailed study of aggregated network traffic using time series analysis techniques. The study is based on three sets of packet traces: 175 short-period WAN traces from the NLANR PMA archive (NLANR), 34 long-period WAN traces from NLANR archive (AUCKLAND), and the four Bellcore LAN and WAN traces (BC). We binned the packets with different bin sizes to produce a set of time series estimating the consumed bandwidth. We studied these series using the following time series techniques: summary statistics, time series structure, the autocorrelation function, the histogram, and the power spectral density. Using a qualitative approach, we developed a classification scheme for the traces using the results of our analyses. We believe that this classification scheme will be helpful for others studying these freely available traces.

We studied the predictability of the traces by choosing representatives of the different classes and then applying a wide variety of linear time series models to them. We found considerable variation in predictability. Some network traffic is essentially white noise while other traffic can be predicted with considerable accuracy. The choice of predictive model is also relatively context-dependent, although autoregressive models tend to do well. Predictability is also affected by the bin size used. As might be expected, it is often the case that predictability increases as bin size grows. However, we also found that in many cases there is a "sweet spot", a degree of aggregation at which predictability is maximized.

Effort sponsored by the National Science Foundation under Grants ANI-0093221, ACI-0112891, and EIA-0130869. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation (NSF).

Keywords: network traffic characterization, network traffic classification, network traffic modeling, network traffic prediction, network monitoring, performance analysis

Network Traffic Analysis, Classification, and Prediction

Technical Report NWU-CS-02-11

Department of Computer Science
Northwestern University

Yi Qiao
yqiao@cs.northwestern.edu

Peter Dinda
pdinda@cs.northwestern.edu

January 13, 2003

Abstract

This paper describes a detailed study of aggregated network traffic using time series analysis techniques. The study is based on three sets of packet traces: 175 short-period WAN traces from the NLANR PMA archive (NLANR), 34 long-period WAN traces from NLANR archive (AUCKLAND), and the four Bellcore LAN and WAN traces (BC). We binned the packets with different bin sizes to produce a set of time series estimating the consumed bandwidth. We studied these series using the following time series techniques: summary statistics, time series structure, the autocorrelation function, the histogram, and the power spectral density. Using a qualitative approach, we developed a classification scheme for the traces using the results of our analyses. We believe that this classification scheme will be helpful for others studying these freely available traces.

We studied the predictability of the traces by choosing representatives of the different classes and then applying a wide variety of linear time series models to them. We found considerable variation in predictability. Some network traffic is essentially white noise while other traffic can be predicted with considerable accuracy. The choice of predictive model is also relatively context-dependent, although autoregressive models tend to do well. Predictability is also affected by the bin size used. As might be expected, it is often the case that predictability increases as bin size grows. However, we also found that in many cases there is a “sweet spot”, a degree of aggregation at which predictability is maximized.

1 Introduction

The characteristics and predictability of network traffic are important in many contexts. In this work, we largely focus on network traffic on assorted routers that connect universities and other organizations attached to the Internet. We characterize this traffic using time series techniques and study its predictability using linear time series models. Our study is based on a collection of 213 packet traces from such routers, which we bin at different granularities to produce estimates of the consumed bandwidth as a function of time. We attempt to answer the following questions:

1. What are relevant summary statistics and how are they related?
2. What are the common distributions?

Effort sponsored by the National Science Foundation under Grants ANI-0093221, ACI-0112891, and EIA-0130869. The NLANR PMA traces are provided to the community by the National Laboratory for Applied Network Research under NSF Cooperative Agreement ANI-9807479. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation (NSF).

3. What are the common autocorrelation structures?
4. How can we classify traces based on 1—3?
5. To what extent can we predict traces and what are appropriate models?
6. How do 1—5 depend on the granularity at which we bin?

Notice that an implicit assumption that we make is that the traces are stationary. There is debate within the networking community over whether this is indeed the case [4, 12]. Many of the traces that we examine are clearly stationary. Others are not.

We addressed these questions in several steps. To start, we collected three sets of packet traces for our study. The first set, which we refer to as the NLANR traces, are a random sample of short traces collected from the NLANR PMA system [9]. The second set, which we refer to as the AUCKLAND traces, are a complete set of long traces collected at the University of Auckland. The third set are the well known Bellcore traces. Using the packet arrival times and sizes, we produced time series representing estimates of the consumed bandwidth using binning with non-overlapping bins of different granularities.

In the next part of our process, we characterized each trace at each granularity using a variety of summary statistics and a qualitative examination of four graphs: the time series plot itself, the autocorrelation function (ACF), the histogram, and the power spectral density (PSD) of the traffic. There is a great deal of variation among the traces. In addition, the three sets of traces (AUCKLAND, NLANR, BC) are clearly distinct.

Based on these distinctions, we next developed classifications for the different sets separately. For the 175 short NLANR traces, we developed a classification hierarchy comprising 12 classes, with each class having qualitative selection criteria based on our characterizations. In the case of the AUCKLAND traces, 34 in number, we developed a classification hierarchy with 8 classes. Because the BC traces are few in number, we simply made each trace a class. Given the classifications, we could now reduce our 218 traces to a set of 24 class representatives.

Using the class representatives, we next studied the predictability of the traces using ten linear time series models, which span the range of known structures. Based on our study of the traces' ACFs, we believed that many of the short NLANR traces would prove to be largely unpredictable while the long AUCKLAND traces would fare much better. Our prediction study supported this belief. We demonstrated that all of the classes of the long traces, all of the BC LAN traces, and a portion of the short NLANR traces could be predicted to some useful extent. We also found some variability in the ideal choice of predictive model. Generally, however, it was important for the model to have an autoregressive component. We found that WAN traffic is generally more predictable than LAN traffic. Bin size greatly influences predictability. As one might expect, in many cases predictability increases with increasing bin size. However, such smoothing often does not monotonically increase predictability. About half of the AUCKLAND traces exhibited a "sweet spot," a degree of smoothing at which predictability is maximized, contradicting the work of an earlier prediction study [11].

We believe that our classification scheme and representative traces will be helpful to ourselves and others in the pursuit of techniques to predict network behavior. Hence we have reported on it in considerable depth, including an appendix. Our prediction study is a first step that we are expanding on. The primary focus of this technical report is classification. A companion technical report [10] describes the prediction study in additional detail. The prediction study includes significantly more results than are presented here.

The rest of the paper is organized as follows: Section 2 briefly introduces the three data sets that we use, and describes how we binned the traces to produce our time series. Section 3 reports on the basic statistics of the traces. Section 4 describes the classification system and the time series analysis underlying it, giving many examples. Section 5 focuses on prediction, evaluating the predictability for a representative from

| Name | Number of | | | Duration | Range of Resolutions |
|----------|------------|---------|---------|----------|------------------------------|
| | Raw Traces | Classes | Studied | | |
| NLANR | 175 | 12 | 39 | 90 s | 1,2,4,...,1024 ms |
| AUCKLAND | 34 | 8 | 34 | 1 d | 0.125, 0.25, 0.5,..., 1024 s |
| BC | 4 | n/a | 4 | 1 h, 1 d | 7.8125 ms to 16 s |

Figure 1: Summary of the trace sets used in the study.

each class, the effectiveness of different predictors, and the influence of different bin sizes on predictability. Section 6 concludes the paper.

2 Traces and binning methodology

In order to generalize about our analysis, characterization, and prediction study, it is important to start from a large and varied dataset and to understand the situations for which the data is representative. Our study is based on three different sets of traces shown in Figure 1. The Appendix lists all 223 traces, their classifications, and where to find them.

The NLANR set consists of short period packet header traces chosen at random from among those collected by the Passive Measurement and Analysis (PMA) project at the National Laboratory for Applied Network Research (NLANR) [9]. The PMA project consists of a growing number of monitors located at aggregation points within high performance networks such as vBNS and Abilene. Each of the traces is approximately 90 seconds long and consists of WAN traffic packet headers from a particular interface at a particular PMA site. We randomly chose 175 NLANR traces provided by 13 different PMA sites. The traces were collected in the period April 02, 2002 to April 08, 2002.

The AUCKLAND set also comes from NLANR’s PMA project. These traces are GPS-synchronized IP packet header traces captured with a DAG3 system at the University of Auckland’s Internet uplink by the WAND research group between February and April 2001. These also represent aggregated WAN traffic, but here the durations for most of the traces are on the order of a whole day (86400 seconds). For the present study, we chose 34 traces, collected from February 20, 2001 to March 10, 2001.

The BC set consists of the well known Bellcore packet traces [8] which are available from the Internet Traffic Archive [1]. There are four traces. Two of them are hour-long captures of packets on a LAN on August 29, 1989 and October 5, 1989, while the other two are day-long captures of WAN traffic to/from Bellcore on October 3, 1989 and October 10, 1989.

Each trace contains packet arrivals, with some specific properties of the packet, such as time stamp, packet length, source and destination IP addresses, etc. Of these, we only consider the time stamp and the length, which are sufficient to produce estimates of the consumed bandwidth over time. We generate a very high resolution view bandwidth by binning the packets into non-overlapping bins of a small size. By dividing the sum of the packet sizes in each bin by the bin size, we obtain a time series that estimates of the instantaneous bandwidth usage. Different bin sizes lead to different granularities. Within limits, the estimates become more accurate as the bin size declines, and the length of the time series increases. The time series becomes smoother as the bin size increases, but there are fewer points in it.

Because the lengths of traces in the three sets are different, we use a different set of bin sizes for each set. A typical NLANR trace is 90 seconds long and thus we bias towards smaller bin sizes. We study bin sizes from 0.001 second to 1.024 second. A typical AUCKLAND trace is about 1 day long and thus we bias towards larger bin sizes, choosing the range of 0.125 second to 1024 seconds. For BC traces, we choose bin sizes between 7.8125 ms and 16 seconds. In each case, the bin sizes are exponentially increasing over the range.

| CCs | Min | Max | Mean | Median | Std | Skewness | Cov | Max/Mean | Min/Mean |
|----------|---------|---------|---------|---------|---------|----------|---------|----------|----------|
| Min | 1.0000 | 0.8204 | 0.9819 | 0.9825 | 0.6186 | -0.2719 | -0.5122 | -0.3508 | 0.7739 |
| Max | 0.8204 | 1.0000 | 0.8872 | 0.8850 | 0.8601 | -0.2260 | -0.6219 | -0.3876 | 0.7800 |
| Mean | 0.9819 | 0.8872 | 1.0000 | 0.9999 | 0.7356 | -0.3380 | -0.6022 | -0.4379 | 0.8093 |
| Median | 0.9825 | 0.8850 | 0.9999 | 1.0000 | 0.7305 | -0.3389 | -0.6033 | -0.4370 | 0.8097 |
| Std | 0.6186 | 0.8601 | 0.7356 | 0.7305 | 1.0000 | -0.4067 | -0.6232 | -0.5670 | 0.6737 |
| Skewness | -0.2719 | -0.2260 | -0.3380 | -0.3389 | -0.4067 | 1.0000 | 0.6643 | 0.8282 | -0.3331 |
| Cov | -0.5122 | -0.6219 | -0.6022 | -0.6033 | -0.6232 | 0.6643 | 1.0000 | 0.8415 | -0.7768 |
| Max/Mean | -0.3508 | -0.3876 | -0.4379 | -0.4370 | -0.5670 | 0.8282 | 0.8415 | 1.0000 | -0.5537 |
| Min/Mean | 0.7739 | 0.7800 | 0.8093 | 0.8097 | 0.6737 | -0.3331 | -0.7768 | -0.5537 | 1.0000 |

Figure 2: Correlation Coefficients between basic statistical properties.

| CCs | Min | Max | Mean | Median | Std | Skewness | Cov | Max/Mean | Min/Mean |
|----------|--------|---------|--------|--------|---------|----------|---------|----------|----------|
| Bin Size | 0.7368 | -0.3981 | 0.8543 | 0.3823 | -0.5452 | -0.5573 | -0.5452 | -0.3981 | 0.7368 |

Figure 3: Correlation Coefficients between bin size and some statistics of a NLANR trace.

3 Summary statistics and their relationships

As a first step, we studied the summary statistics of the traces and how they are related. In the following, we will use the NLANR traces as our example. The results for AUCKLAND traces and BC traces are quite similar.

For each trace and each bin size, we computed the following summary statistics: min, max, mean, median, std, skewness, cov, max/mean, and min/mean. Min, max, mean and median are the minimum, maximum, average, and median value of instantaneous network traffic of the trace, respectively. Std is the standard deviation of network traffic, while cov stands for coefficient of variation, which is the value of standard deviation divided by average network traffic. Skewness represents the skewness of the network traffic distribution. At last, max/mean and min/mean is the ratio of maximum and minimum instantaneous network traffic to average network traffic.

It is interesting to note that there are strong relationships between these statistics. Figure 2 gives the correlation coefficients (CCs) between our nine summary statistics over the 175 NLANR traces using a bin size of 0.008 seconds. For example, the correlation coefficient between std and mean is 0.7356, which means as the average level of traffic in the trace increases, it will also have increasing variability. On the other hand, the CC between mean and cov is -0.6022, indicating that in relative term, variance will decrease as mean traffic increases. The CC of mean and max/mean shows that increasing average traffic volume will cause the ratio of max/mean to decrease.

Another important question is how different bin sizes can influence these statistical properties. For example, how will coefficient of variation change as we smooth the trace? What about minimum and maximum traffic? Figure 3 shows the CCs between bin sizes and the statistics for a typical NLANR trace, ADV-1017869359-1. The CCs between bin size and the max/mean ratio, standard deviation and COV are negative, which means that as we go to finer granularities, we can expect to get greater variation, both in relative and absolute terms. Note that the high value of the mean CC is not significant. The variation in the mean as we increase bin size is tiny in absolute terms.

4 Hierarchical classification of the traces

Our trace classification scheme is based primarily on the visual time series properties of each trace as seen under different bin sizes. For each trace and bin size, we plotted the time series plot, the autocorrelation

function, the histogram, and the power spectral density (PSD). We then studied these plots. For the autocorrelation function, we considered its overall shape, the fraction of the coefficients that are significant, and the strength of the coefficients. For the histogram, we tried to characterize its skew, presence or absence of multimodality, and the tail. For the PSD, we used a log-log plot, making it easy to detect long-range dependence and periodicity. We looked for periodicities and the general slope of the plot.

Using these characterizations of the plots, we classified the traces into a hierarchy. The three sets of traces behaved quite differently and so we developed a separate classification hierarchy for the NLANR short and long period traces. We didn't have enough traces within the BC set to do a meaningful classification, so we just report on the characteristics of each trace.

It is important to realize that while the time series properties we used for classification are objective, our classification scheme is basically subjective. We classify traces based on the *appearance* of various time series analysis plots. However, each of our classes is, on its face, different from the others. Furthermore, we have arranged our classes into a hierarchy. As we move up the hierarchy to more general classes, these classes are increasingly different. The reader can cut through our hierarchy at a level at which he is comfortable.

Using our criteria, we now consider each of the three sets of traces, showing our classification hierarchies and examples for each of the classes. We then give additional details on our hierarchical classification.

4.1 NLANR short period WAN traces

We divided the NLANR traces into 12 classes, which we illustrate graphically in Figure 4 and summarize in Figure 5. Figure 4 shows a hierarchical classification of NLANR traces, where each of the 12 leaf nodes represents a class. Some of the 12 classes also appear as internal nodes, indicating that they merge the behaviors of their descendent classes and describe them in a courser way. For example, we could classify all the NLANR traces into two "superclasses", class 2 and class 4.

Figure 5 summarizes the characteristics of each of the classes, briefly describing its traffic distribution, bandwidth utilization, features of ACF plot and PSD plot, percentage of significant coefficients in the ACF, as well as the proportion of NLANR traces that are in this class.

Next, we'll look at the more common classes. For each class we describe, we will look at one representative trace. We studied each class using 11 different bin sizes: 0.001 seconds, 0.002 second, 0.0004 second, 0.008 second, 0.016 second, 0.032 second, 0.064 second, 0.128 second, 0.256 second, 0.512 second and 1.024 seconds.

4.1.1 Class 1

Figure 6 shows the time series (upper left) of the network traffic for trace ADV-1017901188, a typical class 1 trace, using a bin size of 0.001 second. It also shows the autocorrelation function (upper right), histogram (lower left), and the power spectral density (PSD, lower right) of the traffic.

In the ACF plot, we see that for any lag greater than zero the ACF effectively disappears. In total, only 4% of the coefficients are significant at a significance level (p -value) of 0.05. This suggests that the traffic is basically white noise, and hence strongly suggests that traces in this class will be unpredictable. Although we do see some periodic fluctuations in the ACF, the amplitude of these fluctuations is very small and is likely to have very little impact on the predictability of the signal.

The histogram is a rough estimate for the distribution of the traffic signal. We have found, using a Q-Q plot, that this distribution is heavy-tailed, that it has the form $y = x^{-\alpha}$. The histogram and the time series plot together indicate that there this link is underutilized.

The PSD plot is almost flat, that is, the overall amplitude of the PSD does not change as the frequency increases. This means that the original time series of traffic covers the whole frequency band, containing

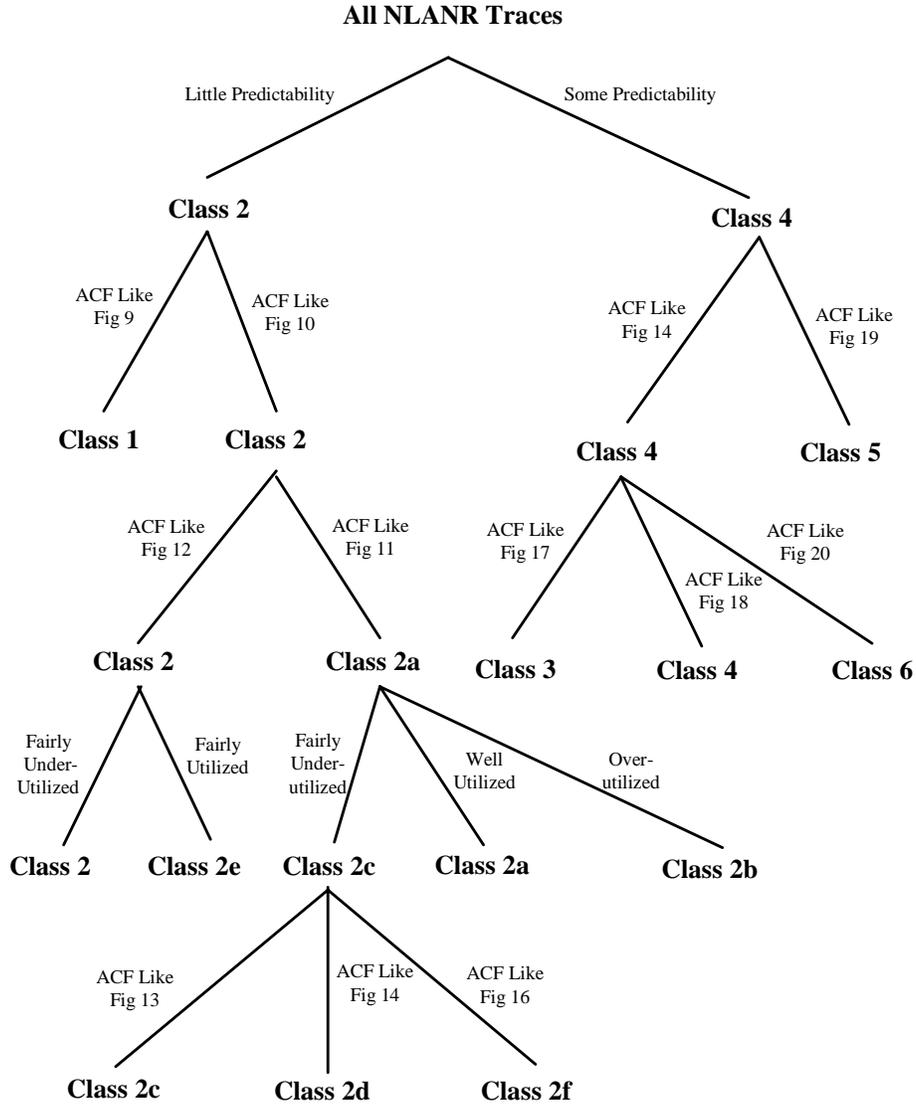


Figure 4: Hierarchical classification of the NLANR traces.

both high-frequency components and low-frequency components, and all of these frequency components are roughly equally important for the time series. This is again indicative of white noise and low predictability. The slight fluctuations in the ACF show up here as slight peaks.

All four plots change as we vary the bin size. These changes are a part of our classification. Using the same trace as before, Figures 7 through 9 show the plots using bin sizes of 0.008 s, 0.128 s and 1.024 s, respectively. The effect of bin size is clearly visible. As the signal is smoothed, the trace slowly becomes more predictable (notice the increase of significant ACF coefficients in Figure 8) and then declines (ACF in Figure 9).

In the remaining classes, we will not show figures for each of the bin sizes. Instead, we will show one set of plots for each class, using the bin size for each of the plots that best emphasizes the observation made in the text. For the present class, we would use the histogram plot at bin size of 0.001 s to best describe the distribution of the network traffic, the ACF plot at bin size 0.008 s to best describe the ACF, the time

| Class Type | Traffic Distribution Description | Bandwidth Utilization | ACF Plot | Significant ACFs | PSD Plot Description | Proportion |
|------------|--|-------------------------|----------|------------------|---|------------|
| 1 | Heavy-tailed distribution in the form of $y = x^{-\alpha}$, Fig 6 | underutilized | Fig 6 | 3%-15% | Flat, Fig 6 | 8.00% |
| 2 | Multiple independent distribution, each with a heavy tail, Fig 10 | Somewhat Unter-utilized | Fig 10 | 3%-20% | Almost Flat, Fig 10 | 15.43% |
| 2a | Combination of half a normal distribution and a heavy-tailed distribution like class 1, Fig 11 | Well Utilized | Fig 11 | 5%-10% | Flat, Fig 11 | 6.86% |
| 2b | Special Distribution, Fig 12 | Over Utilized | Fig 12 | 1%-5% | Flat | 3.43% |
| 2c | Similar as class 2a or 2, Fig 13 | Somewhat Utilized | Fig 13 | 4%-20% | Flat | 20.57% |
| 2d | Similar as class 2 or 1 | Somewhat underutilized | Fig 14 | 1%-3% | Flat | 15.43% |
| 2e | Special distribution, containing multiple independent distributions, Fig 15 | Somewhat Utilized | Fig 15 | 2%-26% | Almost Flat | 6.29% |
| 2f | Similar as class 2 | Somewhat underutilized | Fig 16 | 4%-13% | Flat, with two significant high frequency components | 3.43% |
| 3 | Similar as class 2e | Somewhat Utilized | Fig 17 | 73%-88% | Linearly decrease as frequency increase | 0.57% |
| 4 | Similar as class 2 or 2e | Somewhat Utilized | Fig 18 | 58%-85% | Linearly decreasing, with some significant high frequencies, Fig 18 | 7.43% |
| 5 | Combination of half a normal distribution and a heavy-tailed distribution like class 1, Fig 19 | Well-utilized | Fig 19 | 61%-65% | Some decreasing with a dominant frequency, Fig 19 | 6.86% |
| 6 | Similar as class 2e or 2 | Somewhat Utilized | Fig 20 | 20%-38% | Linearly decrease as frequency increase | 5.71% |

Figure 5: Summary for all classes in NLANR traces.

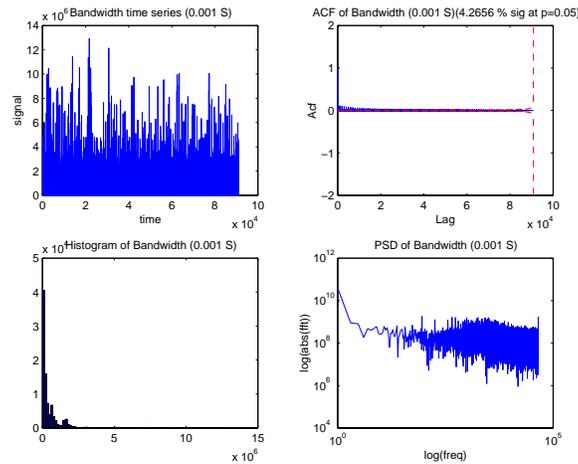


Figure 6: Time-series plot, ACF plot, Histogram of traffic, and PSD plot for a NLANR class 1 trace using bin size 0.001 s.

series plot at 0.008 s to best illustrate the under-utilization of the link, and the PSD plot at 0.128 s to find the frequency component causing the fluctuation in the ACF, and the time series plot at the 1.024 s bin size to best depict how the network traffic has changed over the period of 90 seconds.

4.1.2 Class 2

Figure 10 shows the four plots for trace AIX-1018214469, a typical class 2 trace. The time series plot, the histogram plot and PSD plot all use a bin size of 0.001 s, but the ACF plot uses a 0.128 s bin size. As discussed in the previous section, we have chosen the bin sizes that most clearly enforce the points in the

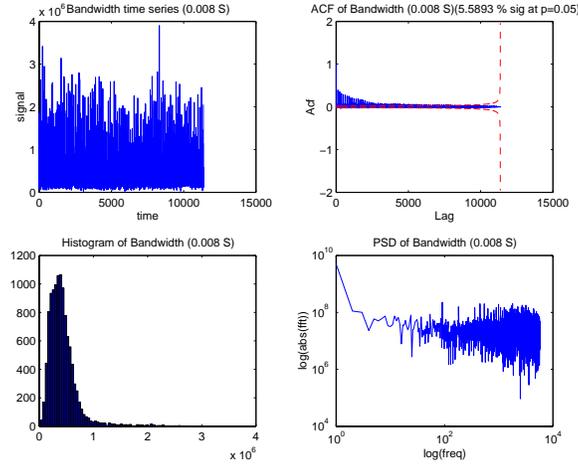


Figure 7: Time-series plot, ACF plot, Histogram of traffic, and PSD for a NLANR class 1 trace using bin size 0.008 s.

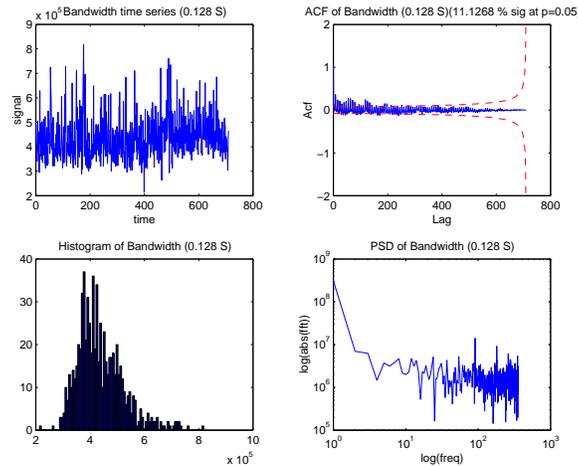


Figure 8: Time-series plot, ACF plot, Histogram of traffic, and PSD plot for a NLANR class 1 trace using bin size 0.128 s.

text. In the discussion of subsequent NLANR classes, we will use a bin size of 0.128 s for the ACF plot and 0.001 s for other three plots.

The ACF plot shows that 20% of the coefficients are significant, which is much higher than that of the class 1 trace. However, the amplitude of these coefficients is quite small. The ACF curve demonstrates a weak but observable low frequency behavior: it gradually goes from a positive value to zero at some lag on x-axis, and continue to decrease until reaching its minimum, then begins to gradually increase.

Like class 1, the PSD plot is still quite flat as a whole, but we can also find a small, linear decline for the first half of the PSD curve: the higher the frequency, the lower the amplitude, meaning that the lower frequency components of the signal are slightly more significant than higher frequency components. This is the reason why we can see some low-frequency-wise behavior in the ACF plot. Regrettably, the significance of low-frequency components is weak, so traces in this class will still have little predictability, if any. In fact, for all other bin sizes of this trace, the percentage of significant ACFs are even lower than 20%. We can only clearly observe the low-frequency-wise behavior of ACF plot under bin sizes around 0.1 second.

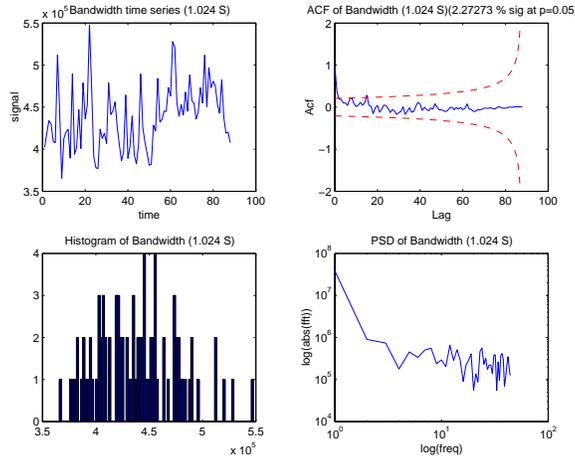


Figure 9: Time-series plot, ACF plot, Histogram of traffic, and PSD plot for a NLANR class 1 trace using bin size 1.024 s.

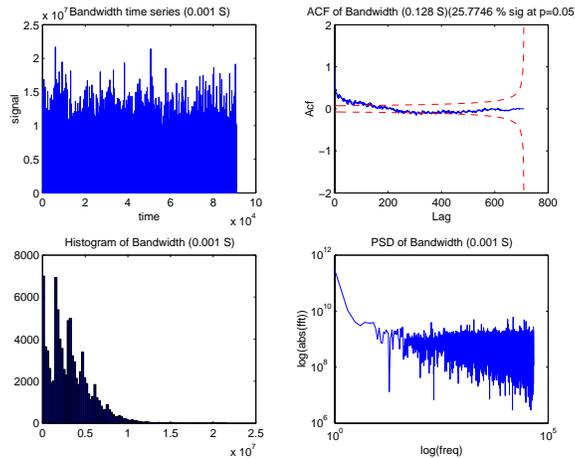


Figure 10: Time-series plot, ACF plot, Histogram of traffic, and PSD plot for a NLANR class 2 trace.

Similar to class 1, there are also some high-frequency fluctuations in the ACF plot for the trace. Although the fluctuations here are much weaker, they still indicates the existence of some significant high-frequency component, which can be directly observed from PSD plot.

The histogram of the trace is multimodal, although with a heavy tail. One explanation is that the trace shows several very deterministic flows at particular bandwidths. Between the histogram and the time series, it is clear that the link is underutilized.

As shown in Figure 4, class 1 and class 2 can be merged into a “superclass” of classes that have little hope for predictability. The major difference between traces in class 1 and class 2 lies in their ACF plots (compare Figure 7 (class 1) and Figure 10 (class 2)).

4.1.3 Class 2a

Figure 11 shows trace MRA-1018106960, a representative of class 2a. Here we use 0.128 s for the ACF plot and 0.001 s for the remaining plots.

Only about 5% of the coefficients in the ACF plot are significant, and there are only very weak and

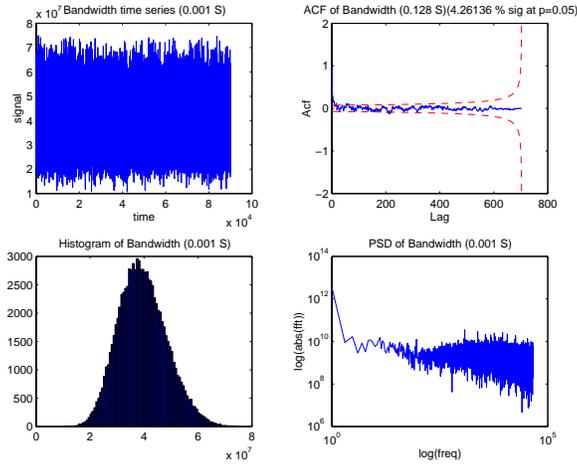


Figure 11: Time-series plot, ACF plot, Histogram of traffic, and PSD plot for a NLANR class 2a trace.

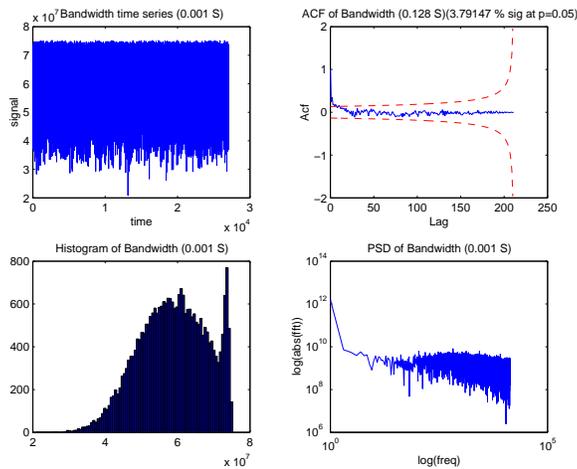


Figure 12: Time-series plot, ACF plot, Histogram of traffic, and PSD plot for a NLANR class 2b trace.

irregular fluctuations in the ACF plot. In addition, the PSD plot is basically flat for the whole frequency band. Therefore, we can make a judgment that traces in this class have limited predictability.

The histogram of the trace is significantly different from what we have seen before. For one thing, it looks symmetric. Using QQ-plots, we have concluded that the left half is very much a normal distribution, while the right half is normal-like, albeit with a heavy tail. This, and the time series plot, suggests to us that the bandwidth on the link is well utilized. It is rarely the case that there is not something to send and it is rarely the case that the link is overcommitted.

4.1.4 Class 2b

Figure 12 shows the four plots for trace MRA-1018149859, a representative class 2b trace. We use 0.128 s for the ACF plot and 0.001 s for the remaining plots.

The ACF plot for the trace shares many common characteristics with that of class 2a, but with even weaker fluctuations and even lower percentage of significant coefficients. This kind of behavior, together with the flat PSD plot, again implies that traces in this class will be difficult to predict.

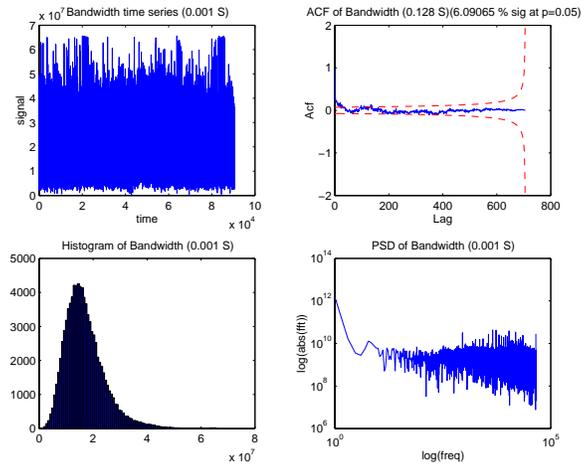


Figure 13: Time-series plot, ACF plot, Histogram of traffic, and PSD plot for a NLANR class 2c trace.

The most significant different between this class and class 2a is the bandwidth utilization. From the time-series plot of Figure 12, we can clearly see that the instantaneous traffic often reaches the bandwidth limit of the link, therefore, bandwidth of this trace is over-utilized. As a consequence, traffic distribution has two separate peaks as seen in the histogram plot, showing its multimodality.

4.1.5 Class 2c

The four plots for the representative class 2c trace, IND-1018043716, are shown in Figure 13. We use 0.128 s for the ACF plot and 0.001 s for the remaining plots.

The traces in this class are very close to traces in class 2a, but there are several subtle differences:

- The bandwidth utilization of the class 2c trace is reasonable, but not as high as that of class 2a. We can see this by the skew in the distribution. The average network traffic of class 2c is lower, and the consumed bandwidth varies less.
- For some class 2c traces, the distribution is more like that of class 2 than class 2a, which contains multiple independent heavy-tailed distributions.
- The low-amplitude fluctuations in the ACF plots for class 2c traces have higher frequency than those in class 2a ACF plots. This leads to slightly higher percentage of significant coefficients, but still suggests little predictability.

4.1.6 Class 2d

Trace ANL-1018064471 is a representative trace for class 2d traces, shown in Figure 14. We use 0.128 s for the ACF plot and 0.001 s for the remaining plots.

This trace has very few significant ACF coefficients, and even weaker fluctuations in the ACF plot than class 2b traces. The PSD plot is still relatively flat.

The histogram in Figure 14 is composed of many peaks, and the distribution appears to have a heavy tail, both characteristics similar to those of class 2. This suggests that the bandwidth of the link is underutilized.

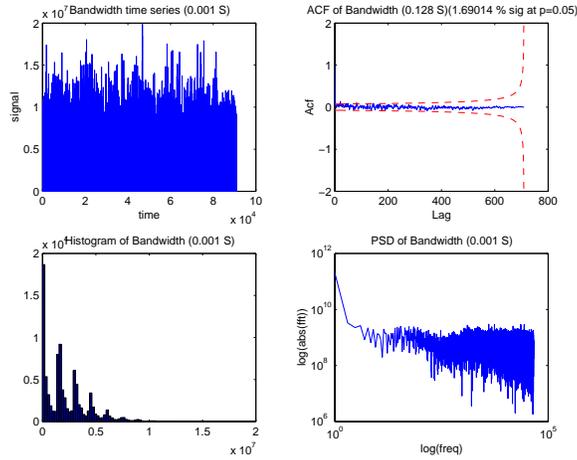


Figure 14: Time-series plot, ACF plot, Histogram of traffic, and PSD plot for a NLANR class 2d trace.

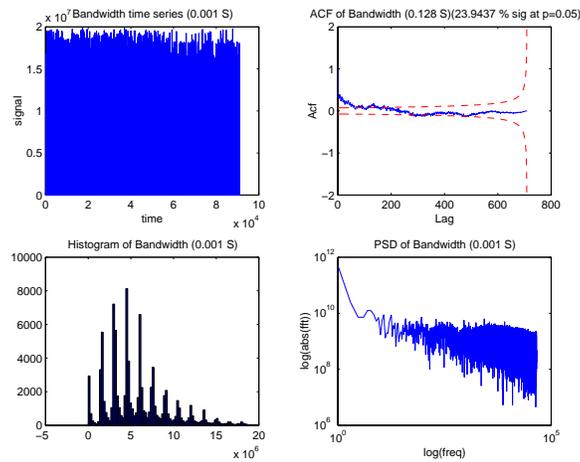


Figure 15: Time-series plot, ACF plot, Histogram of traffic, and PSD plot for a NLANR class 2e trace.

4.1.7 Class 2e

The representative trace for class 2e is BUF-1017869359, with its four plots shown in Figure 15. As before, we use 0.128 s for the ACF plot and 0.001 s for the remaining plots.

The ACF and PSD plots for traces in this class are quite similar to those in class 2. The ACF curve also demonstrates a weak but observable low frequency behavior, together with high frequency local fluctuations. About 24% of the ACF's coefficients are significant. The PSD plot is almost flat, but the first half shows a weak but observable linear decline. However, just like class 2, the downward slope is not strong, nor are there significant frequency components. As a result, traces in class 2e are unlikely to be very predictable.

The histogram is rather unique. Like class 2, there are many modes, suggesting a number of very deterministic flows. A QQ-plot indicates a heavy tail on the distribution. The bandwidth utilization on the link is considerable, higher than in class 2.

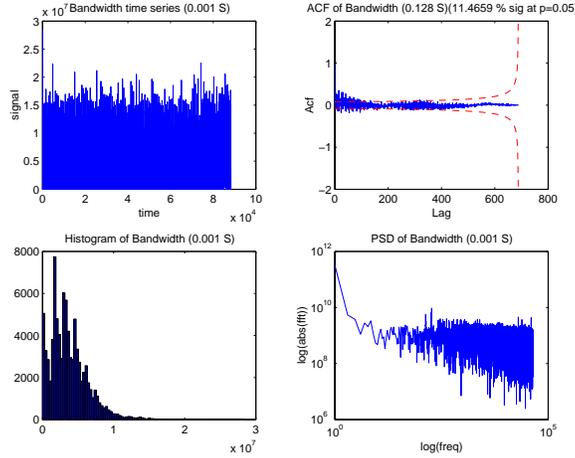


Figure 16: Time-series plot, ACF plot, Histogram of traffic, and PSD plot for a NLANR class 2f trace.

4.1.8 Class 2f

Figure 16 gives the four plot for a representative class 2f trace, TAU-1018150308. As before, we use 0.128 s for the ACF plot and 0.001 s for the remaining plots.

Similar to the other classes we have discussed to this point, the traces in this class have a small percentage of significant ACF coefficients (11% in the figure) and a relatively flat PSD plot, suggesting limited predictability.

However, here the PSD is not purely flat at all of the frequencies, instead, we can clearly see two dominant high frequency components. Correspondingly, the fluctuations in the ACF curve have a higher frequency and larger amplitudes than those in all the other classes.

Ignoring the left-most peak in the plot, the histogram for class 2f is very close to those for class 2: it is multimodal and has a heavy tail. The link is somewhat underutilized.

Returning to the classification hierarchy of Figure 4, we have noted the similarity of class 2c, class 2d and class 2f: they all are likely to have limited predictability due to their ACF characteristics, they all exhibit somewhat underutilized bandwidth, their ACF plots share common features with that of class 2a. These classes are differentiated due to subtle differences in their ACF plots.

4.1.9 Class 3

The class 3 traces are radically different from what we've seen before. The four plots for a representative trace, ANL-1018225955, are shown in Figure 17. We use 0.128 s for the ACF plot and 0.001 s for the remaining plots.

Obviously, the ACF plot for this class is quite different from what we have seen up to this point. Over 70% of the coefficients in ACF plot are not only significant but also large. We can clearly see a quite strong low-frequency-wise behavior. This low-frequency behavior is much stronger than that in the class 2 traces, and the whole curve is much smoother. For smaller bin sizes, the percentage of significant coefficients is even higher, often over 85%. The PSD plot shows the dominant low frequencies more clearly: as the frequency increases, the PSD amplitude is not flat at all; on the contrary, it drops quickly and linearly in this log-log plot. The linearly decrease of the PSD plot is actually a good indication of long-range dependence of the signal. It is highly likely that traces in this class will exhibit predictability.

Similar to class 2e, the histogram is multimodal, suggesting several very deterministic flows. The envelope of the distribution looks like the combination of half a normal distribution as the left branch and a

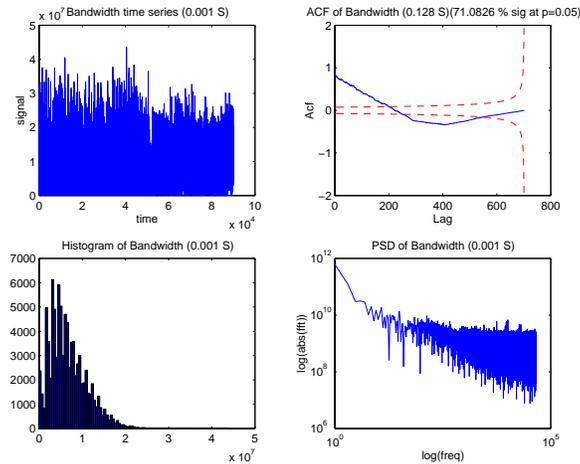


Figure 17: Time-series plot, ACF plot, Histogram of traffic, and PSD plot for a NLANR class 3 trace.

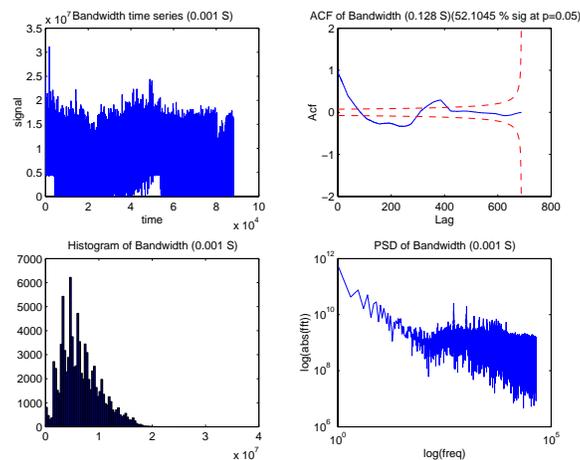


Figure 18: Time-series plot, ACF plot, Histogram of traffic, and PSD plot for a NLANR class 4 trace.

heavy tailed distribution as the right branch. The link bandwidth is somewhat utilized.

4.1.10 Class 4

The four plots for a class 4 trace, APN-1018150307, are shown in Figure 18. We use 0.128 s for the ACF plot and 0.001 for the other plots.

Traces in this class are similar to those in class 3. The histogram is similar to either class 2 or class 3, and indicates a fairly well utilized link. The linear decrease in the PSD plot, together with over 50% significant coefficients in ACF plot suggest predictability.

However, unlike the smooth ACF curve in class 3, here the ACF plot contains very high frequency fluctuations, making the curve thicker than usual. Therefore, besides the dominating low-frequency components in the signal, there are also some significant high-frequency components, which can also be readily seen on the PSD plot.

The trace is also clearly nonstationary, as can be seen by the behavior of the time series graph. There are epochs of different behavior.

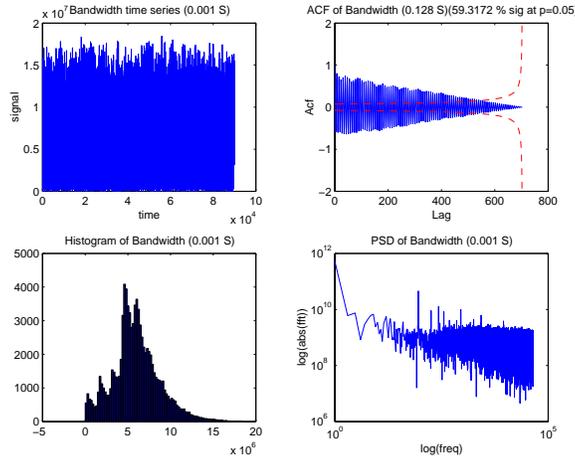


Figure 19: Time-series plot, ACF plot, Histogram of traffic, and PSD plot for a NLANR class 5 trace.

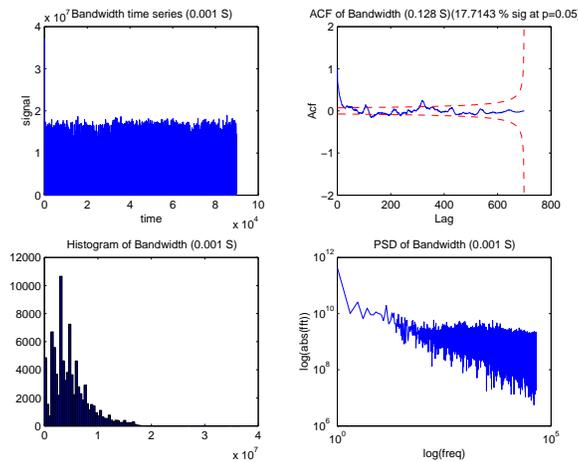


Figure 20: Time-series plot, ACF plot, Histogram of traffic, and PSD plot for a NLANR class 6 trace.

4.1.11 Class 5

Figure 19 shows a representative trace of class 5, COS-1018021588. As before, we are using bin size of 0.001 s normally and 0.128 s for the ACF plot.

The ACF plot for this trace is quite interesting: the coefficients quickly oscillate across the x-axis at a high frequency and a large amplitude. We saw similar behavior in class 2, but the amplitude here is much higher. About 55% of the coefficients are significant and large for a bin size of 0.128 s, and the percentage grows as we look at smaller bin sizes. The PSD plot shows us a narrow peak that is responsible for our ACF oscillations. Based on these observations, we expect that the traces in this class will be quite predictable.

The envelope of the histogram is again the combination of half a normal distribution and a heavy-tailed distribution in the form of $y = x^{-\alpha}$. We can also see several small independent components in the histogram. The bandwidth utilization of this link is high.

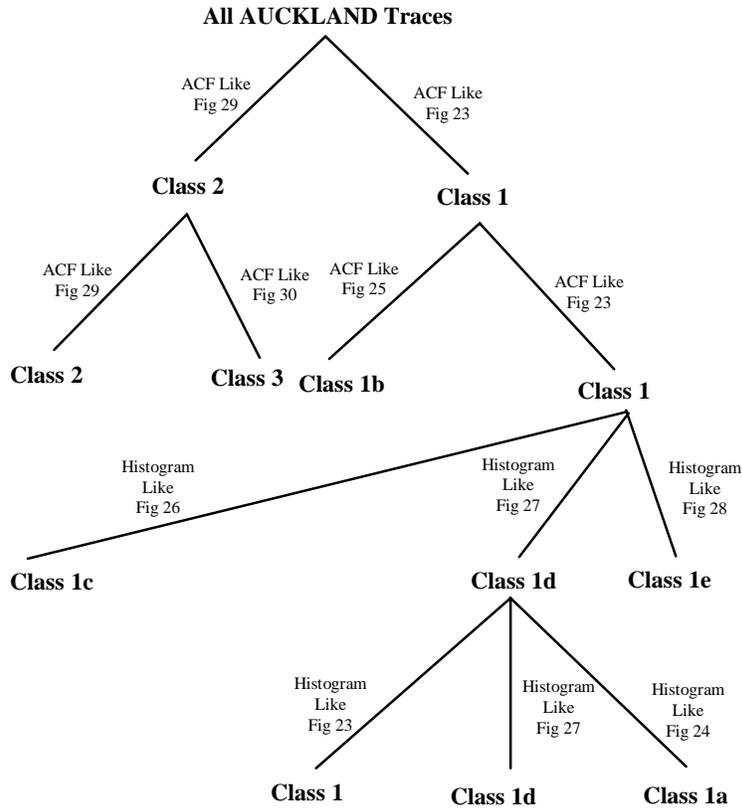


Figure 21: Hierarchical classification for AUCKLAND traces.

4.1.12 Class 6

Figure 20 is for trace OSU-1017717808, a representative class 6 trace. As before, we are using bin size of 0.001 s normally and 0.128 s for the ACF plot.

The ACF plot looks rather different from the other traces we have seen up to this point. The curve is composed of some irregular fluctuations with small amplitudes and a seemingly periodical pulse. Although only about 18% of the coefficients are significant, the PSD plot does slowly decrease with frequency, indicating that low frequencies dominate. We would expect less predictability classes 3 to 5, but larger than the other NLANR classes.

The histogram for a class 6 trace is either almost the same as that of the class 2e trace, or quite close to that of the class 2 trace. Bandwidth utilization for class 6 is fair.

4.2 AUCKLAND long period WAN traces

For classification of the AUCKLAND traces, we adopted different bin sizes: 0.125 s, 0.25 s, 0.5 s, 1 s, 2 s, 4 s, 8 s, 16 s, 32 s, 64 s, 128 s, 256 s, 512 s, and 1024 s. We looked at coarser grain bins because the traces are very long and provide enough data to do so. We eliminated smaller bins because the cost of computing our various figures proved to be prohibitive. We divide the AUCKLAND traces into 8 classes, as illustrated graphically in Figure 21 and summarized in Figure 22. In the following we will describe the classes, using one representative trace for each.

| Class Type Type | Traffic Distribution Description | Bandwidth Utilization | ACF Plot | Significant ACFs | PSD Plot Description | Proportion |
|-----------------|--|-----------------------|----------|------------------|---|------------|
| 1 | Contains two overlapping parts: a heavy tailed distribution in the form of $y = x^{-\alpha}$, and a distribution similar as that of class 2, Fig 23 | Well Utilized | Fig 23 | 88%-95% | Linearly decrease as frequency increase, Fig 23 | 30.30% |
| 1a | Similar as class 1 except the two parts are more overlapping, Fig 24 | Somewhat Utilized | Fig 24 | 71%-90% | Linearly decreasing, Fig 24 | 21.21% |
| 1b | Roughly a heavy-tailed distribution in the form of $y = x^{-\alpha}$, Fig 25 | Somewhat Utilized | Fig 25 | 73%-91% | Linearly decreasing | 15.15% |
| 1c | Multiple heavy-tailed distribution, Fig 26 | Somewhat Utilized | Fig 26 | 87%-95% | Linearly decreasing | 3.03% |
| 1d | A special distribution containing two parts Fig 27 | Well Utilized | Fig 27 | 76%-95% | Linearly decreasing | 3.03% |
| 1e | Lognormal distribution, Fig 28 | Well Utilized | Fig 28 | 86%-94% | Linearly decreasing | 9.09% |
| 2 | Combination of half a normal distribution and a heavy-tailed distribution, Fig 29 | Somewhat Utilized | Fig 29 | 60%-78% | Linearly decreasing, Fig 29 | 12.12% |
| 3 | Combination of half a normal distribution and a heavy-tailed distribution, Fig 30 | Well Utilized | Fig 30 | 45%-55% | Linearly decreasing | 6.06% |

Figure 22: Summary for all classes in AUCKLAND traces.

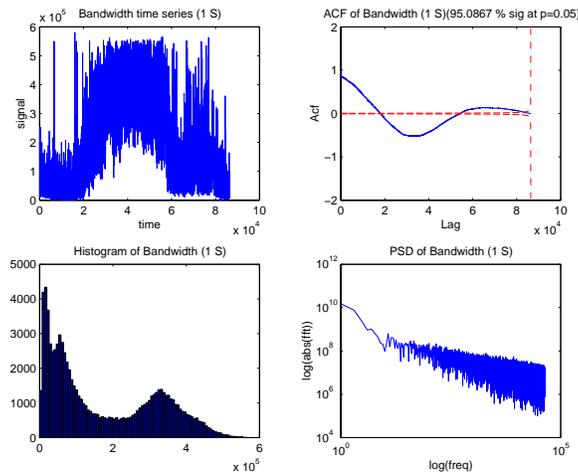


Figure 23: Time-series plot, ACF plot, Histogram of traffic, and PSD plot for an Auckland class 1 trace.

4.2.1 Class 1

Figure 23 gives the four plots for trace 20020221-020000-0, a typical class 1 trace. Note that the time series plot, ACF plot, the histogram plot and PSD plot all use a bin size of 1 s. For the other AUCKLAND classes that we illustrate, we will do the same.

From the first point of view, it is obvious that each of these plots looks very different from the plots in short NLANR traces. The ACF plot is very smooth and regular, it oscillates across the x-axis with large amplitude. The percentage of significant coefficients is quite high: 95% of the coefficients are significant and strong for this trace at a bin size 1 s.

The PSD plot is not flat as in most of the short NLANR traces. Instead, in this log-log plot, the curve decreases nearly linearly as the frequency increases, showing strong dominance of low frequency components in the signal. More importantly, the linear decrease of spectral density implies that this trace shows long-range dependence. In addition, we could also explain the smoothness of the ACF curve by observing that there are no particular significant high frequency components in the PSD plot. We can be reasonably sure that traces in this class will be quite predictable.

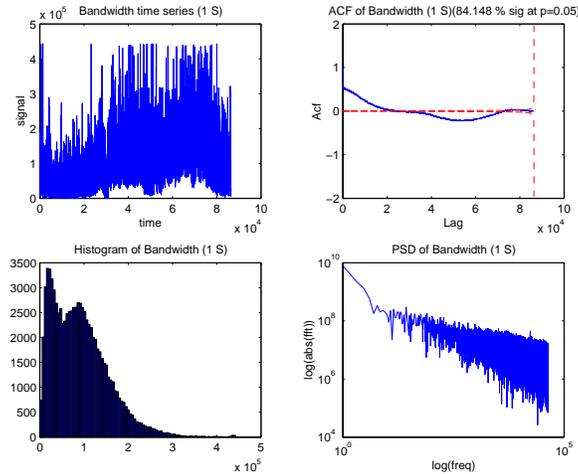


Figure 24: Time-series plot, ACF plot, Histogram of traffic, and PSD plot for an Auckland class 1a trace.

The trace’s histogram is complex. The core of distribution is heavy tailed ($x^{-\alpha}$) and there is a “lump” suggesting a secondary element to the distribution. The two elements are overlapping, but their peaks are still far apart. This is a time-of-day effect and can be explained by considering the time series plot of the traffic: there is a period of relatively low bandwidth usage, followed by a high usage period, and then followed by a low usage period. Knowing that the duration of the trace is exactly one day, we can easily understand this behavior. Note that this makes the trace is nonstationary. The bandwidth utilization of the link in this class is good.

4.2.2 Class 1a

The plots of a trace in class 1a, 20010225-020000-1, are shown in Figure 24, all of which are based on bin size of 1 s.

This class has only some minor differences compared to class 1. The ACF plot still oscillates around the x-axis slowly (85% of coefficients are significant and strong), but the curve is not as regular as the ACF plot for class 1. The PSD is still decreasing linearly in the log-log plot as the frequency increases, thus indicating long-range dependence and good predictability. The histogram is still bimodal as in class 1, except that here the peaks are much closer together. The time series plot still exhibits time-of-day effects, indicating nonstationarity, but the effects are not as obvious as with class 1. The link bandwidth utilization for this class is fair.

4.2.3 Class 1b

Figure 25 gives the four plots for a representative class 1b trace, 20010224-020000-1. As before, 1 s is the bin size.

Although the ACF plot for this trace is not radically different from those in class 1 and class 1a, there are differences. The coefficients are smaller and the curve is rougher. However, over 90% of the coefficients are significant and we again see a linearly decreasing PSD plot, suggesting both long-range dependence and a high predictability.

Although this trace also is one day long, the gap between the high and low traffic periods is smaller. As a result, we no longer see multimodality in the histogram. Instead, we claim that the histogram can be

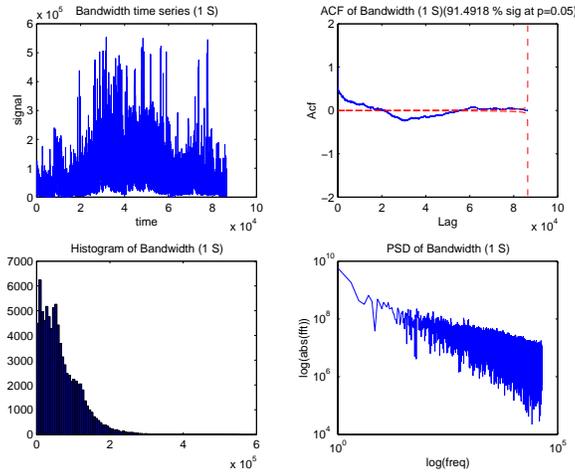


Figure 25: Time-series plot, ACF plot, Histogram of traffic, and PSD plot for an Auckland class 1b trace.

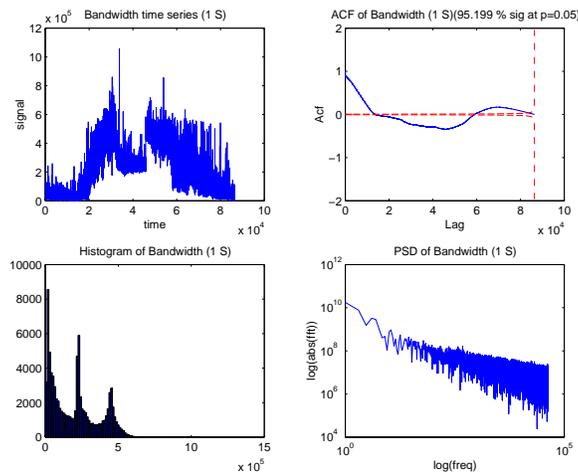


Figure 26: Time-series plot, ACF plot, Histogram of traffic, and PSD plot for an Auckland class 1c trace.

approximated with merely a single, heavy-tailed distribution. The bandwidth utilization for this class is as high as class 1 and class 1a.

4.2.4 Class 1c

Figure 26, which shows the four plots for 20010308-020000-0, a typical class 1c trace. As before 1 s is the bin size.

Except for the histogram, this class is almost the same as class 1a: the ACF plot is smooth but not as regular as that of class 1, about 95% of the coefficients are significant, and the PSD plot is linearly decreasing, indicating long-range dependence and likely good predictability.

The histogram is trimodal with one peak for each of the three obvious periods that can be seen in the time series plot. Clearly, this is a nonstationary trace. The link bandwidth is somewhat utilized.

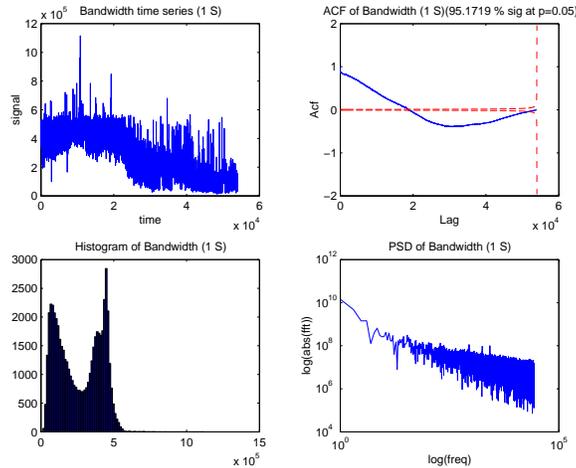


Figure 27: Time-series plot, ACF plot, Histogram of traffic, and PSD plot for an Auckland class 1d trace.

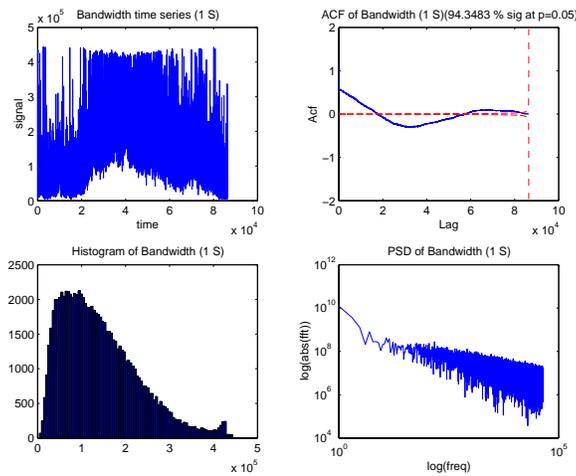


Figure 28: Time-series plot, ACF plot, Histogram of traffic, and PSD plot for an Auckland class 1e trace.

4.2.5 Class 1d

The four plots for the representative class 1d trace, 20010301-110023-0, are shown in Figure 27, again using a 1 s bin size.

Here we see a smooth and regular ACF curve, with over 95% of the coefficients being significant and most being large. The PSD plot is linearly decreasing, suggesting long-range dependence. Both of these features are almost the same as those of class 1 and suggest good predictability.

The histogram consists of two peaks. For about the first half of the trace, the link operated at or near capacity, while during the second half, it operated well below capacity.

4.2.6 Class 1e

Trace 20010221-020000-1 is a typical trace of class 1e. Its four plots are depicted in Figure 28 using 1 s bins.

This class has ACF and PSD plots very similar to those of class 1. About 94% of the coefficients in this

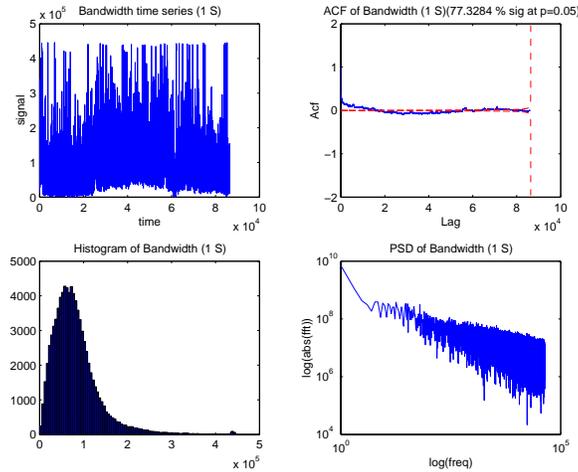


Figure 29: Time-series plot, ACF plot, Histogram of traffic, and PSD plot for an Auckland class 2 trace.

trace’s ACF are significant. We conclude that there is long-range dependence and a high likelihood that the trace is very predictable.

Like class 1b, we could not find multimodality in the histogram. The distribution can be approximated by a lognormal distribution. Bandwidth is well utilized for this class.

4.2.7 Class 2

Figure 29 gives the four plots for a trace of class 2, 20010224-020000-1, at a bin size of 1 s.

The PSD plot here indicates long-range dependence for the trace. For ACF plot, 77% of the coefficients are significant. Although even here there is a slow oscillation across the x-axis of the ACF plot, the amplitude is very low compared to classes 1 and 1a—the ACF coefficients here are much weaker than the ones we’ve seen earlier. We can also see higher frequency fluctuations on the ACF curve, which makes it rougher. We conclude that there is probably some predictability in this class, but it will be much weaker than the other AUCKLAND classes.

The heavy tail of the histogram is very obvious here. The bulk of the distribution looks normal, however. The bandwidth utilization for the trace is fair, but due to the heavy tail, the distribution is obviously skewed. For the time series plot, the gap between high bandwidth usage period and low bandwidth usage period is far smaller than that of class 1. We could arguably claim that this trace is stationary.

4.2.8 Class 3

The last class in the AUCKLAND traces is class 3. The four plots of a representative trace, 20010301-110023-1, are shown in Figure 30. A 1 s bin size is used.

For both ACF plot and PSD plots, this class is similar to class 2 except for some details in the curves. The percentage of significant ACF coefficients is even lower, only about 55% in the representative trace. Like class 2, class 3 will have some degree of predictability, but the predictability will not be as good as class 1 through class 1e.

The histograms of class 3 are similar to those of class 2. However, the average network traffic is much larger. The link bandwidth is well-utilized.

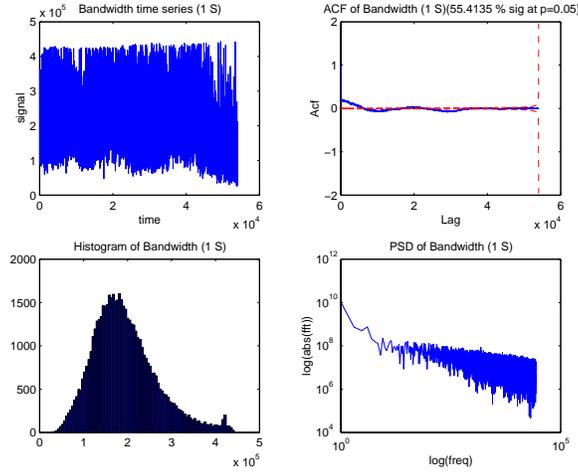


Figure 30: Time-series plot, ACF plot, Histogram of traffic, and PSD plot for an Auckland class 3 trace.

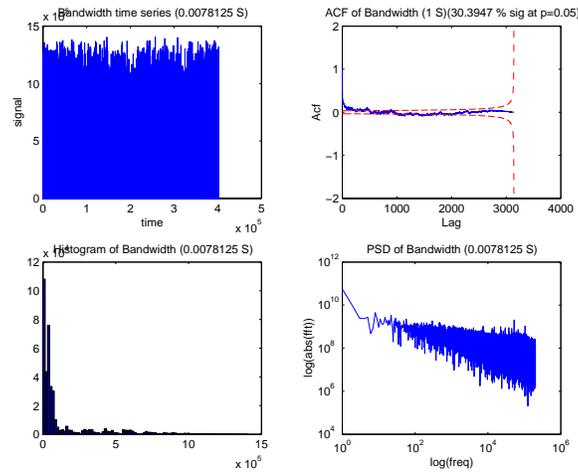


Figure 31: Time-series plot, ACF plot, Histogram of traffic, and PSD plot for trace BC-pAug89.

4.3 BC (LAN) Ethernet traces

We have only 4 BC traces to study, which is an insufficient number to do any sort of hierarchical classification. Two of the traces are of WAN traffic while the other two are of LAN traffic. Since we have thus far discussed a large number of current WAN traces, we change pace here and describe the two BC LAN traces in detail.

4.3.1 Trace BC-pAug89

The four plots for this trace are shown in Figure 31. All the plots are using a bin size of 0.0078125S except for the ACF plot, which is at bin size of 1 s for visual clarity. The total length of the trace is about 3120 seconds.

About 30% of the ACF coefficients are significant. Smaller bin sizes result in a larger percentage of significant coefficients. However, the coefficients remain weak. The log-log PSD plot is decreasing linearly with increasing frequency. Of course, this is the key long-range dependence discovery described in the

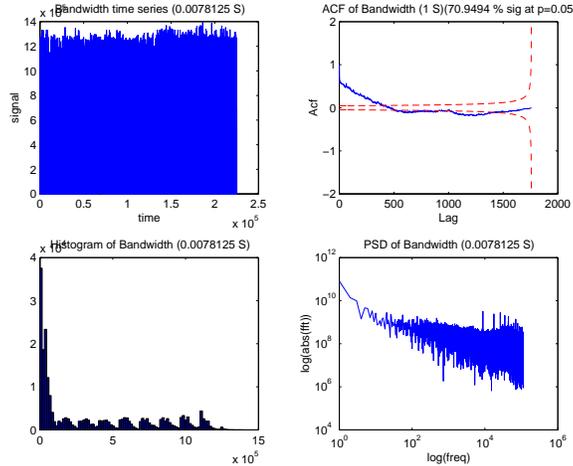


Figure 32: Time-series plot, ACF plot, Histogram of traffic, and PSD plot for trace BC-pOct89.

original paper on these traces. Predictability is possible, but will probably be limited due to the weakness of the coefficients. The histogram of traffic is heavy-tailed. The link utilization is low.

4.3.2 Trace BC-pOct89

The plots for this trace are shown in Figure 32. As before, all the plots are using a bin size of 0.0078125S except for the ACF plot, which uses 1 s bins for visual clarity. The trace is about 1760 seconds long.

About 71% of the ACF coefficients are significant, and more become significant at smaller bin sizes. The coefficients are stronger than in the other BC LAN trace. The PSD plot decreases linearly, again suggesting the long-range dependence property that these traces are famous for. The histogram is heavy tailed, but there are clearly a number of modes, possibly suggesting several deterministic flows. The link bandwidth utilization is low.

4.4 Using the hierarchical classification scheme

We have now described our classes for both the NLANR traces and the AUCKLAND traces. This section describes how to use the classification hierarchies that we have derived (i.e., the trees and tables of Figures 4, 5, 21, and 22.) The difference between different classes is often large, but can also be very subtle. Sometimes subtle differences are important, but at other times, we only care about large differences. Our hierarchical classification scheme takes into account these considerations.

Figure 4 shows the classification tree for the NLANR traces, while Figure 21 shows the tree for the AUCKLAND traces. The Appendix lists each of the traces and their class.

Let's use Figure 4 as an example. Each leaf node represents a NLANR trace class with the corresponding class name in the node. Each internal node is also labeled with, but here with the name of the most typical class among its children. For example, the three leaf nodes at the bottom of the figure are marked with "class 2c", "class 2d", "class 2f", respectively. Since these are all leaf nodes, each of them stands for a separate class. When we aggregate these classes together, we choose the class that is most common among the children, "class 2c", as the name of the parent class. To repeat, the class of each internal node is simply the most typical class within the subtree below it. The leaf nodes represent the "core" classes in our classification scheme and are what are described in Figures 5 and 22, and used in the list of classified traces in the Appendix.

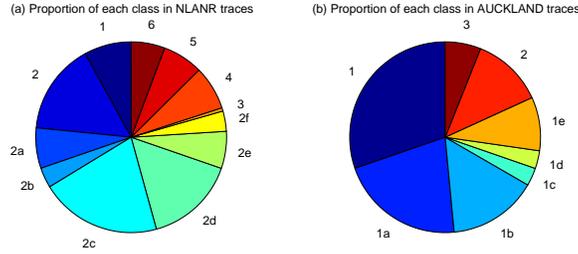


Figure 33: Proportion of different classes for NLANR traces and AUCKLAND traces.

Our classification scheme can be mainly used in two ways. First, for any two classes, we can easily see how different the classes are from each other, and what these differences are. For example, we can see from Figure 4 that class 2a and class 2b (the two leaf nodes) are very close to each other, that they all have good predictability, and that their ACF plots are also quite similar, the only difference being that bandwidth of class 2a is well utilized, while bandwidth for class 2b is overutilized.

Our classification scheme can also be used to classify traces at different granularities. For example, if we only need to split the NLANR traces into two classes, we can just look at the two children of the root node, each child node stands for such a class. In this hypothetical classification, we'll name the class of the left child node class A, the class of the right child node class B. Class A will include the traces in class 1 to class 2f of our final classification, and class B will include class 3 to class 6. The leftmost node, marked with "class 2", indicates that that class 2 is a typical subclass for class A. Similarly, class 4 is a typical subclass for class B. For each trace, we either put it into class A, or put it into class B according to the percentage of significant coefficients in its ACF plot. If we wish to break class B to subclasses, then we come to its two children, marked with "class 4" and "class 5", respectively. As shown in the figure, this time, the classification is based on the appearance of the ACF curve. Recursively, we can go to an even finer granularity, up to the final 12 classes.

4.5 Summary

To view the trace classification at the higher level, let's look at Figure 5 and Figure 22, which has summarized the characteristics for all the different classes in NLANR traces and AUCKLAND traces respectively. Figure 33 (a) and (b) shows the proportion of each of the classes in the NLANR and AUCKLAND traces, respectively.

We have found that most of the short-period NLANR traces probably have limited predictability. They tend to have a small percentages of significant coefficients in their ACFs and their significant coefficients tend to be very small. The long-period AUCKLAND traces present the opposite story. Interestingly, all of the traces that appear to have a useful degree of predictability as gauged by their ACFs also demonstrate long-range-dependence. The BC LAN traces show limited prospects for predictability.

5 Traffic prediction

Given our classified traces, we now consider how predictable the traces in each class are with linear time series models. Linear models attempt to capture the autocorrelation function in an efficient manner. As such, a trace whose autocorrelation function is essentially zero is unlikely to be predictable using these models. However, by studying their predictability directly, we can gain further insights and quantify the degree of predictability. This is especially interesting for traces that are clearly nonstationary and hence are not well modeled by their autocorrelation functions.

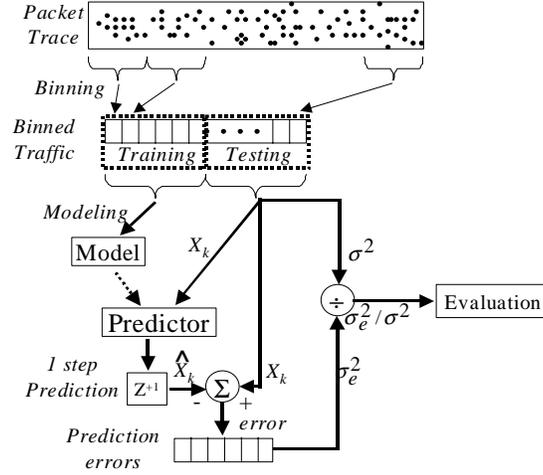


Figure 34: Methodology of binning prediction.

5.1 Evaluation criterion and prediction models

Figure 34 illustrates the methodology and the discussion in this section. To evaluate predictability of a trace, we use the first half of the binned trace as the training sequence for the prediction model, and the second half as the testing sequence for the predictor that we generate from the model. The model contains fitted parameters that approximate the autocorrelation function of the first half of the binned trace. The predictor contains these parameters and state, particularly observations of the testing sequence and of its own prediction errors.

We denote an actual value in the testing sequence as x_i , while the one-step-ahead prediction for that value, using all data in the training and testing sequence up to and including x_{i-1} , is denoted as \hat{x}_i . The prediction error is then

$$e_i = \hat{x}_i - x_i$$

Characterizing the error sequence $\langle e_i \rangle$ for all i in the testing sequence is what we mean by evaluating the predictability. We use the following three criteria and their accompanying metrics:

- Does the predictor provide gain over simply using the mean value of the sequence? For a completely independent sequence (white noise), the mean value is the best possible predictor. We measure this using the ratio of mean squared error ($\sigma_e^2 = \sum_i e_i^2$) to the variance of testing sequence ($\sigma^2 = \sum_i (\mu - x_i)^2$), that is, an estimate of

$$\frac{E[(\hat{x}_i - x_i)^2]}{E[(E[x_i] - x_i)^2]}$$

μ is the average over the testing sequence. This is the *noise-to-signal ratio* of the predictor running on the trace. If the ratio is less than one, then it means that the predictor is doing better than using the average. The smaller the ratio is, the more benefit there is to using the predictor.

- Have we extracted all the predictability? If we have done so, then the error sequence should be white noise. We test for white noise by looking at the percentage of coefficients in the autocorrelation function of the errors is significant. We call this metric *sigacfrac*. The smaller this value is, the more likely it is that the errors are independent.

- Are the errors normally distributed? If the errors are not only independent, but also normally distributed, then computing confidence bounds on functions of multiple predictions becomes much easier. Hence, this is a desirable property. To test it, we do a QQ-plot of the empirical error distribution and a normal distribution. We then compute the R^2 value for a least-squares fit to this plot. We use that value, which we refer to as *r2normfit*, as our metric. The closer this value is to 1, the more likely it is that the error distribution is normal.

In addition to the above metrics, our toolset produces (and we have recorded), a wide variety of other metrics. We compute all of our metrics for combination of trace, bin size, and prediction model.

There are many possible prediction models that could be used for traffic prediction. In this paper, we use a variety of prediction models provided by the RPS toolkit.¹ These models span all the commonly used linear models. The predictors we use are

- MEAN: MEAN uses the long-term mean of the signal as a prediction. Its noise-to-signal ratio is typically 1.0 for obvious reasons. Its `sigacfrac` and *r2normfit* are identical to those of the testing sequence itself.
- LAST: LAST simply uses the last observed value as the prediction for the next value.
- BM(32): BM(32) predicts that the next value will be the average of some window of up to the 32 previous values. The size of the window is chosen to provide the best fit to the training sequence.
- MA(8): MA(8) is a moving average model of order 8. It computes the next value as a linear function of its previous 8 errors.
- AR(8) and AR(32): These are autoregressive models of order 8 and 32, respectively. They compute the next value as a linear function of the last 8 or 32 observed values.
- ARMA(4,4): This model combines 4 autoregressive parameters and 4 moving average parameters. The prediction is a linear combination of the previous 4 errors and observations.
- ARIMA(4,1,4) and ARIMA(4,2,4): These are once and twice integrated ARMA(4,4) models. They difference the signal once or twice, run an ARMA(4,4), and once or twice integrate its predictions. Unlike the previous models, they can capture a simple form of nonstationarity.
- ARFIMA(4,-1,4): This model is a “fractionally integrated” ARMA model that can capture the long-range dependence of self-similar signals. In some sense, these signals are borderline-stationary.

The AR, MA, ARMA, and ARIMA models are classical time series models described by Box, et al [3]. ARFIMA models are well covered in more recent literature [7, 6, 2]. The RPS technical report [5] also provides an explanation of these models as well as a detailed description of the implementations we use here. The same implementations are used for offline and online analysis in RPS.

It is important to point out that we do *not* follow the Box-Jenkins methodology in choosing our models. The methodology involves a human being in the loop, making it difficult to do offline with large numbers of datasets and impossible to do online within a resource prediction system. Furthermore, the methodology, and measures such as AIC, do not take into account the run-time of the predictors or of model fitting. We have chosen our models to make a reasonable first exploration of the space of linear models applied to network prediction. Our models cover all of the different classes of models and their orders are chosen to be as large (and hence descriptive) as possible while still remaining practical in a real system. Running

¹RPS is available from <http://www.cs.northwestern.edu/~RPS>.

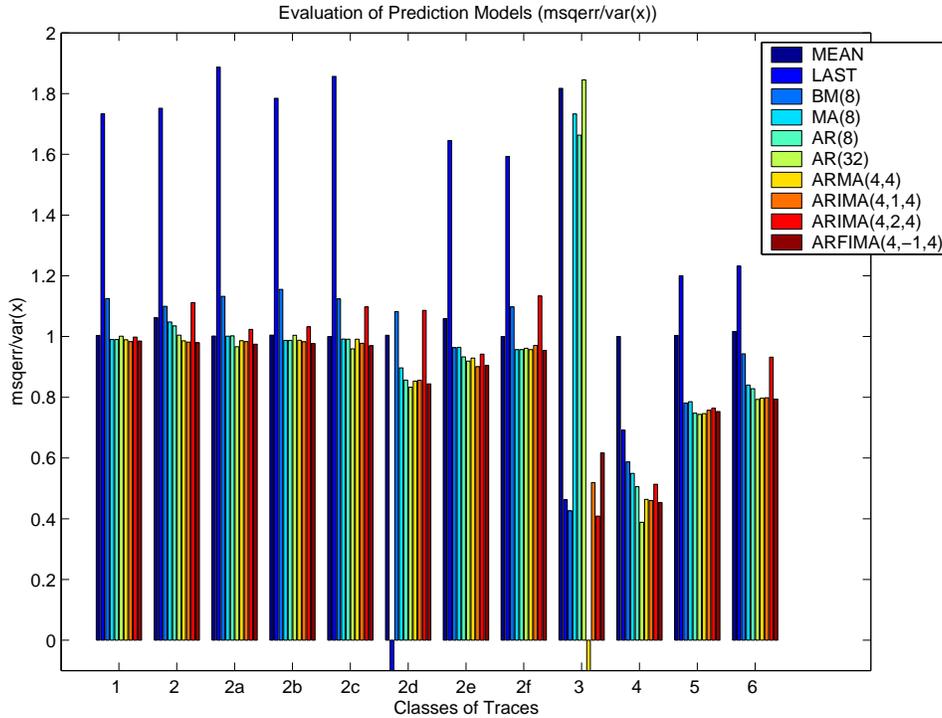


Figure 35: Prediction performance in terms of noise-to-signal ratio using different prediction models on representative traces of all the NLANR classes. Smaller is better.

different models on a trace lets us plumb the predictability of the trace itself and compare the performance of different models at extracting it.

To exclude the potential influence of bin sizes on predictability, we use one particular bin size for each trace when comparing different prediction models. This bin size is 0.008 seconds for the NLANR traces, and 8 seconds for the AUCKLAND traces. For the BC traces, we use 0.125 seconds for the two LAN traces and 8 seconds for the two WAN traces. For each trace we studied, we also evaluated the influence of different bin sizes, i.e., the granularity of the traffic data, on the degree of the predictability. In other words, we measured the relationship between different granularities and predictability of a trace. In this, we focused on the AR(32) model.

The primary focus of this technical report is classification. A companion technical report [10] describes the prediction study in additional detail. The prediction study includes significantly more results than are presented here.

5.2 Predicting the NLANR traces

Figures 35 through 37 show the performance of the prediction models on representatives of the different classes of NLANR traces. Figure 35 shows the noise-to-signal ratio, Figure 36 shows *sigacfrac*, and Figure 37 shows *r2normfit*. All the predictions here are using bin size of 0.008 second.

Using these figures, we can not only evaluate the effectiveness of different prediction models, but also gauge the predictability of the different classes of traces. As we mentioned earlier, our classification schemes for both NLANR traces and AUCKLAND traces are closely related, using properties that are good indicators of the predictability of a trace, such as significant coefficients in the ACF plot, the relationship between amplitude and frequency in the PSD plot, and so on. After running actual predictors, we can consider the

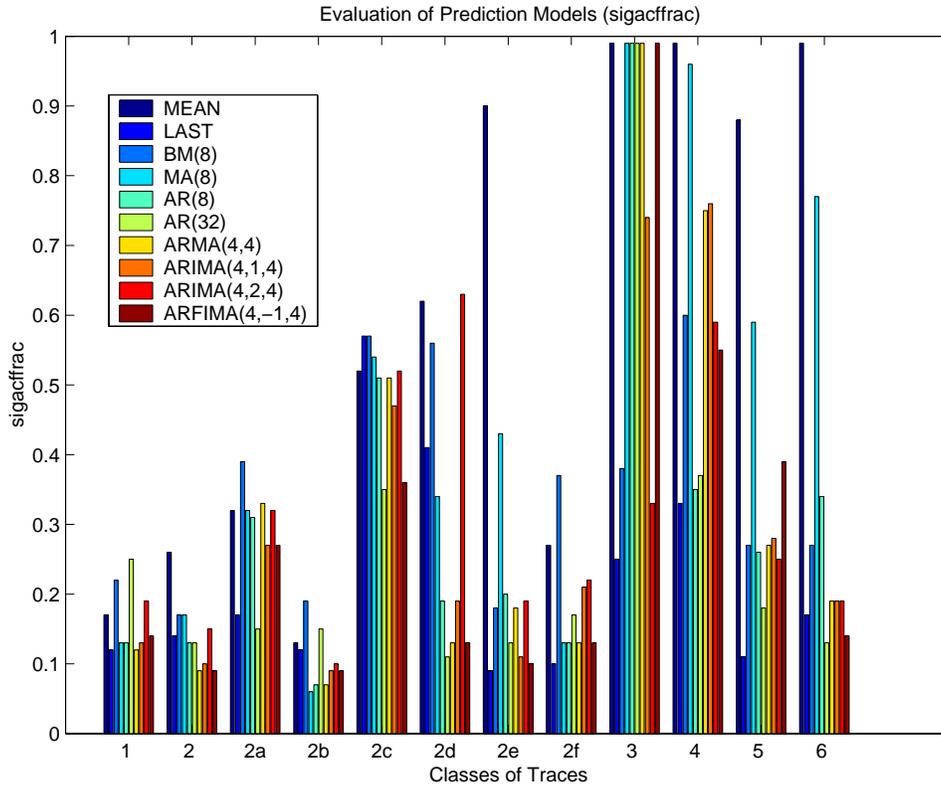


Figure 36: Prediction performance in terms of *sigacfrac* using different prediction models on representative traces of all the NLANR classes. Smaller is better.

effectiveness of these classifications, as well as gain a straightforward understanding of the predictability of each class of traces.

We found the following:

- There are clear differences in the performance of the predictive models with respect to the noise-to-signal ratio (Figure 35). In general, AR(8), AR(32), ARMA(4,4), ARIMA(4,1,4) and ARFIMA(4,-1,4) perform better than other prediction models such as MEAN, LAST, and BM. It usually helps to include an autoregressive component to the prediction. Fractional models (ARFIMA(4,-1,4)) also do quite well. Of these models, the cheapest are the pure autoregressive models. The negative values in the figure indicate combinations that produced either very high error levels or where a predictor failed.
- The actual prediction results for the different classes of traces confirm our expectations of their predictability. Recall that in our classification we divided the NLANR classes into two broad categories. Class 1 through class 2f are the first category, traces that appear to have limited predictability due to their weak ACFs. Class 3 to class 6 are the second category because they appear to have some predictability due to their stronger ACFs. Figure 35 confirms our expectations: For classes 1 through 2f, the noise to signal ratio is generally one, and worse than one for certain models. We are hence convinced that these classes are truly not predictable using linear models. On the other hand, for classes 3 through 6, the best predictor has a much lower ratio, ranging from 0.38 to 0.79. This convinces us that these traces do indeed have linear predictability.

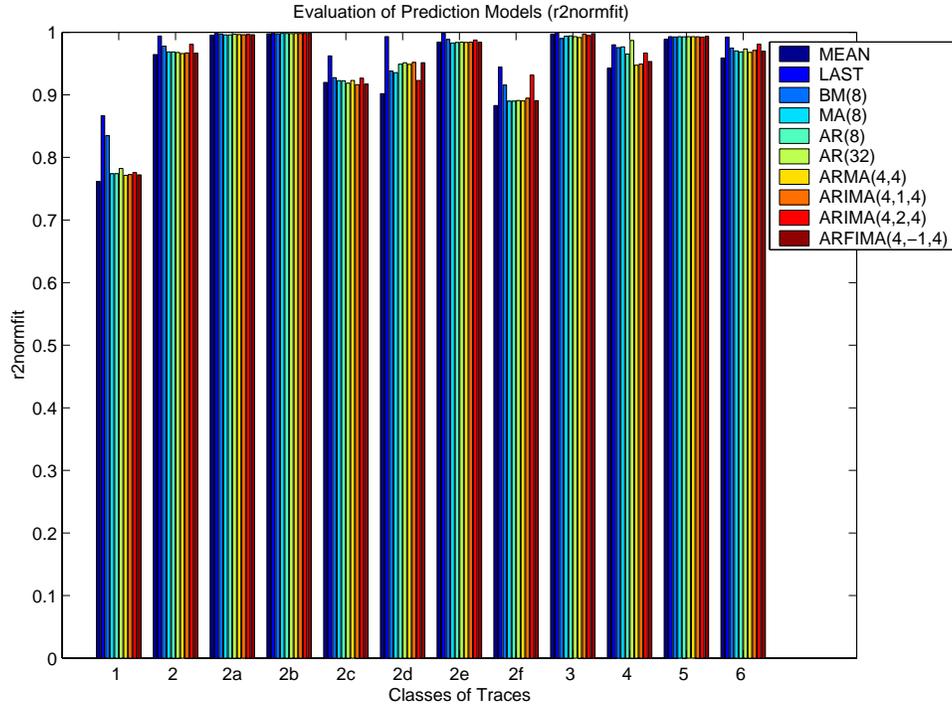


Figure 37: Prediction performance in terms of $r2normfit$ using different prediction models on representative traces of all the NLANR classes. The ideal value is 1.0.

- For almost all the classes and all the predictors, the value of $r2normfit$, our second evaluation criterion for traffic prediction, is very close to 1, meaning that the error distribution are very close to normal distribution, which is a good property that we want.
- Generally, $sigacfrac$, our measure of independence of the prediction errors, tends to be smaller (more independence) for prediction models which contain autoregressive components. Another interesting thing to note here is that, for those classes of traces that have some predictability, the $sigacfrac$ values of their predictors are often larger than the $sigacfrac$ values for non-predictable classes. In other words, for these predictable traces, prediction errors at different times are often correlated with each other, the models are not squeezing out all the possible predictability of the traces.
- About 80% of the traces belong to class 1 through class 2f, which means only a small portion of NLANR traces have some predictability.

Next, we'll consider the influence of bin size on prediction. Recall that binning is how we generated discrete-time signals from the packet traces. The result is an estimate of the instantaneous bandwidth usage that becomes more accurate as the bin size declines. Suppose we first used a large bin size, say, 1.024 second for the NLANR traces, in our prediction evaluation. This would be essentially equivalent to getting the information from the trace using a low pass filter. When we use finer bin sizes, such as 0.008 seconds, we have raised the cut-off frequency of the filter. In other words, when we move from coarser bins to fine bins, we get increasingly high resolution views of the underlying data.

Figure 38 shows the noise-to-signal ratio for the predictable NLANR traces as a function of bin size, ranging from 0.001 s to 1.024 s, doubling at each step. For all predictions here, we use the AR(32) prediction model. Clearly, predictability is a function of bin size, and not just in the simplistic way one might expect.

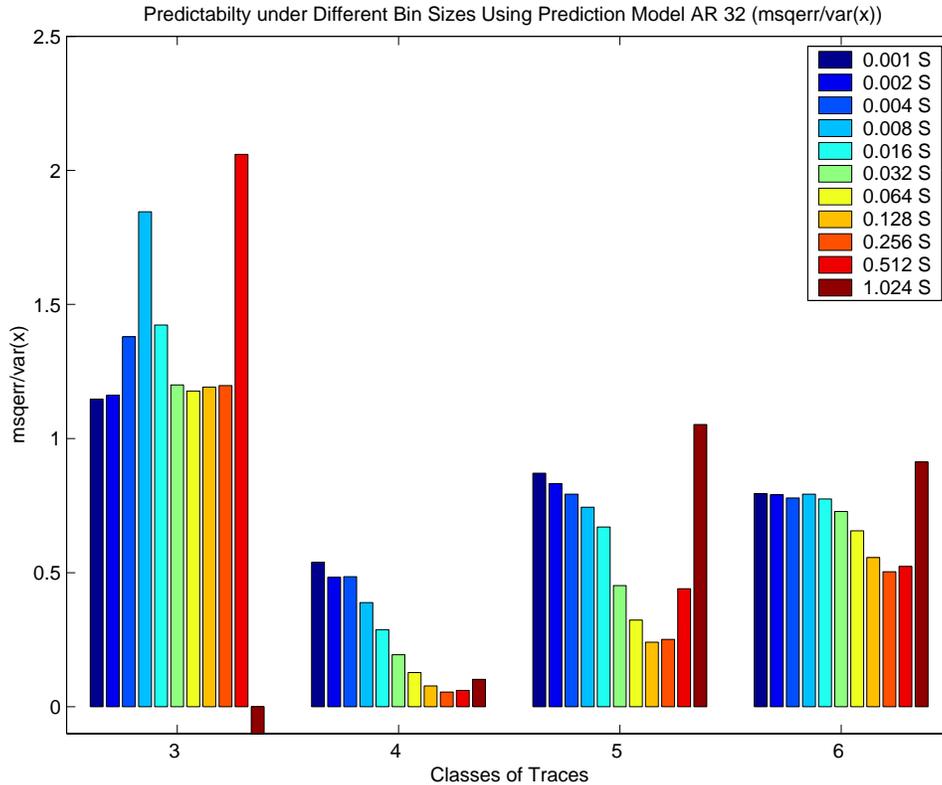


Figure 38: Noise-to-signal performance for NLANR traces as a function of bin size. AR(32) model. Only predictable classes from Figure 35 are shown.

It is not the case that predictability increases monotonically as the trace is smoothed with larger and larger bin sizes.

The most interesting feature here is that the graph shows some concavity for all the four traces: we can clearly see a “sweet spot” for the traffic prediction, where the noise-to-signal ratio is minimized. There is a range of bin size around approximately 0.128 seconds in which the trace is most predictable. The location of the sweet spot varies slightly from trace to trace. For the 4 traces in Figure 38, the sweet spots range from 0.064 second to 0.256 second. For most of the predictable NLANR traces in our study, there exists such a sweet spot.

Further study needs to be done to explain the existence of sweet spot. However, we do have a theory. Consider the behavior of class 5. For this trace, the noise-to-signal ratio at a 1.024 second bin size is very high, slightly over 1. As we gradually decrease bin size toward 0.128 second, the optimal bin size, we are capturing increasingly higher frequency components in the underlying traffic. It may be these components that the predictor is in fact able to model. When we reach 0.128 seconds, this effect is maximized, and we see a ratio of 0.2408. As we further decrease the bin size, more and more high frequency information will be provided to the predictor. It may be the case that this information cannot be modeled by the predictor (i.e., it is noise). As the predictor tries to model it, it is wasting effort that could be spent on the lower frequencies. Hence, the quality of predictions declines. When the bin size finally becomes 0.001 second, the ratio is 0.7945, more than three times larger than with the optimal bin size.

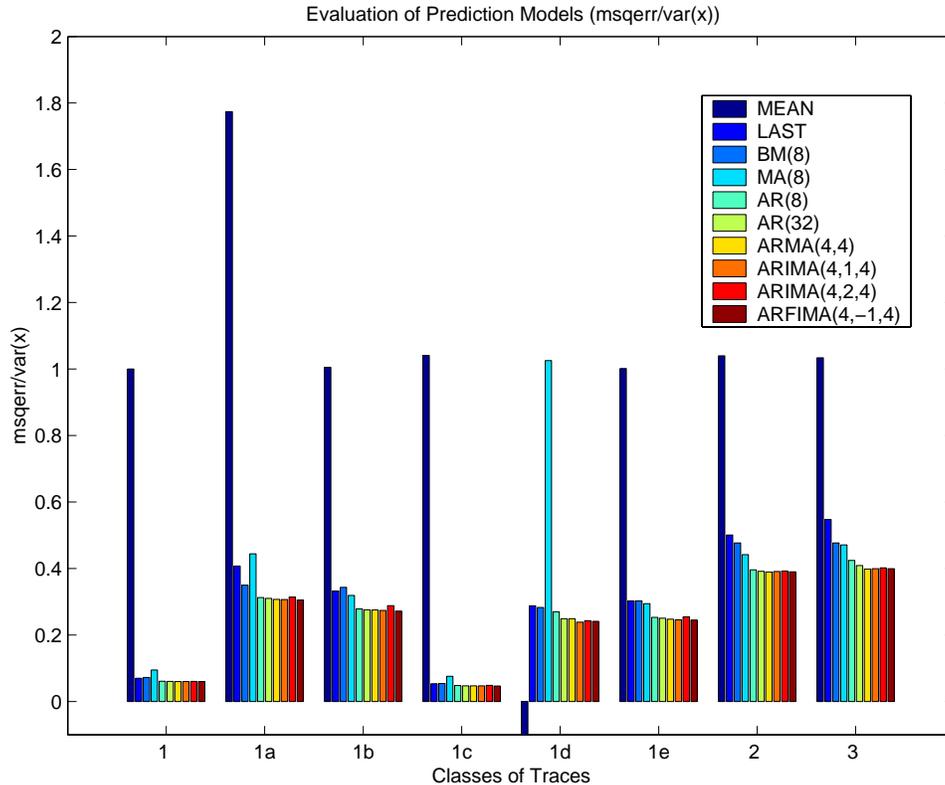


Figure 39: Prediction performance in terms of the noise-to-signal ratio using different prediction models on representative traces of all the AUCKLAND classes. Smaller is better

5.3 Predicting the AUCKLAND traces

Figures 39 through 41 show the performance of the ten predictors on different classes of the AUCKLAND traces. Figure 39 shows the noise-to-signal metric, Figure 40 shows *sigacfrac*, and Figure 41 shows the *r2normfit* metric. The bin size is 8 seconds.

We found the following:

- For the AUCKLAND traces, there is still a considerable variation of performance among the predictors. In almost all cases, LAST, BM, and MA predictors will perform considerably worse, indicating the necessity of including an autoregressive component in the prediction model. It is often helpful to also have a moving average component and an integration. Fractional ARIMA models are effective but do not seem to warrant their high cost for prediction.
- All of these traces are predictable. This can be seen clearly from Figure 39. To be more specific, class 1 to class 1e have very good predictability, the noise-to-signal ratio for the best predictor being between 0.04 to 0.27. Class 2 and class 3 also have some predictability, with the ratio of the best predictor being about 0.40. If we look back to Figure 22, we can find that our expectations of the predictability of each class are confirmed by the measurements here.
- Like short traces, the value of *r2normfit* is very high, so the error distribution can be considered to be normal.
- Except for the MEAN and MA predictors, the values of *sigacfrac* for all other models are quite low.

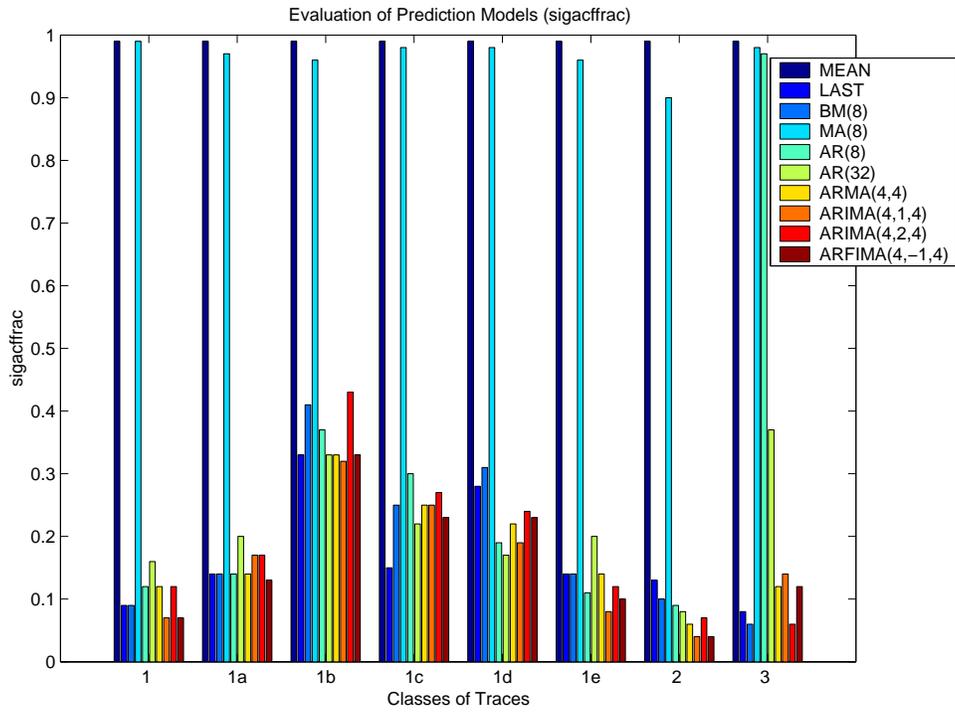


Figure 40: Prediction performance in terms of *sigacfrac* using different prediction models on representative traces of all the AUCKLAND classes. Smaller is better.

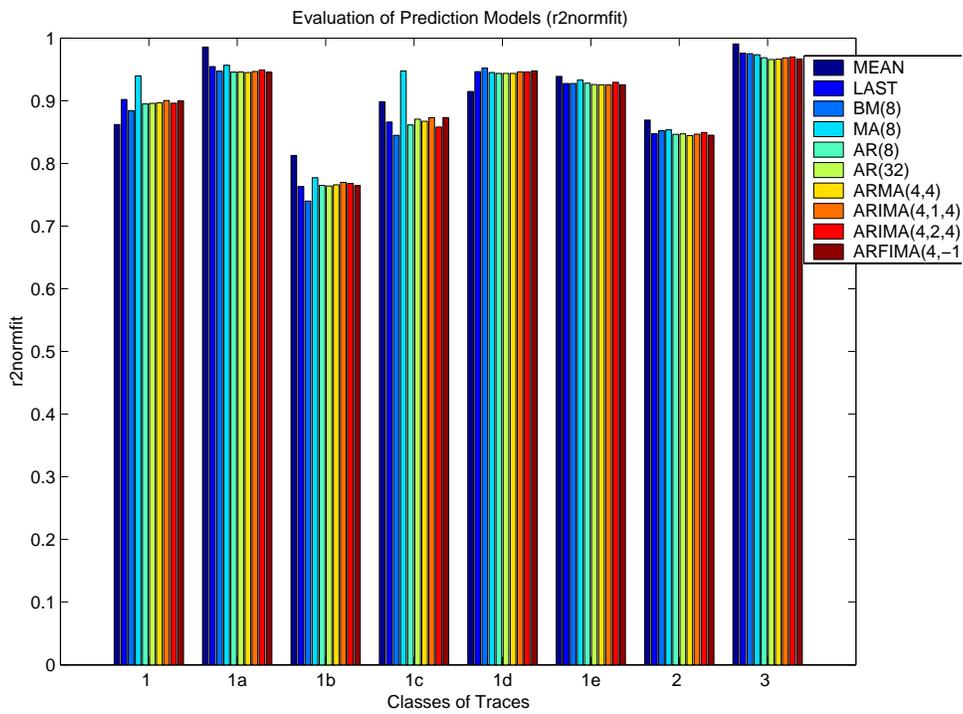


Figure 41: Prediction performance in terms of *r2normfit* using different prediction models on representative traces of all the AUCKLAND classes. The ideal value is 1.0.

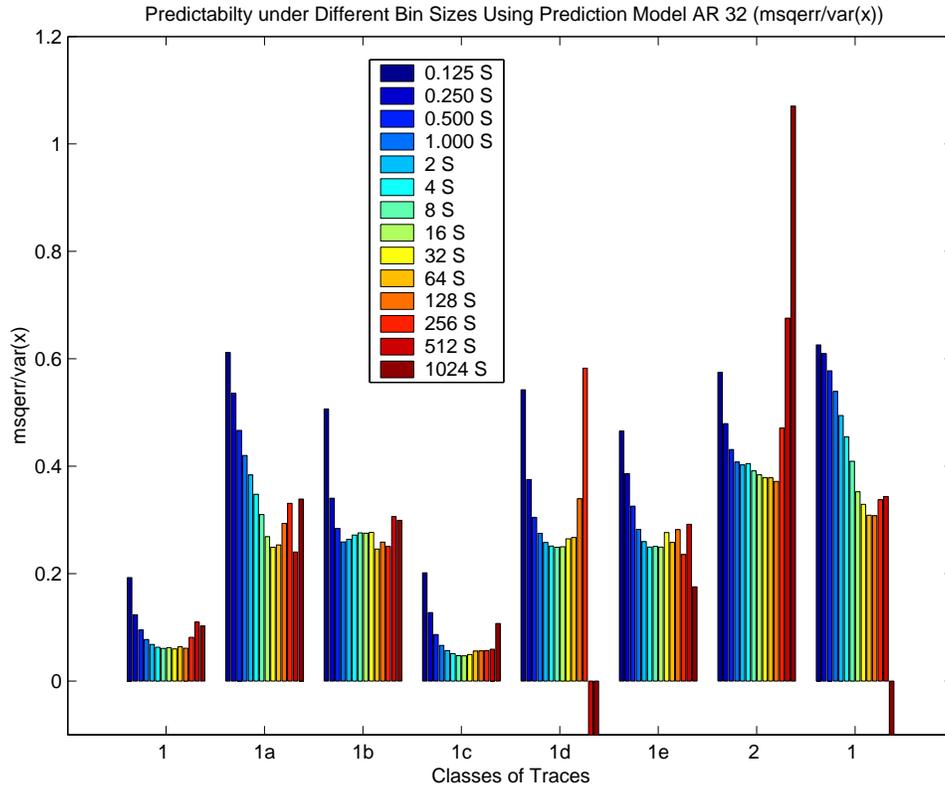


Figure 42: Noise-to-signal performance for AUCKLAND traces as a function of bin size.

In other words, for these predictors, the prediction errors at different times are not closely related, which is a desirable property in a predictor.

As with the NLANR traces, we looked at the effect of bin size. Figure 42 shows the noise-to-signal ratio as a function of bin size ranging from 0.125 s to 1024 s and doubling. The relationship is shown for each of the classes. The AR(32) predictor is used. Again, we can clearly see that there is a relationship between prediction quality and bin size.

Like the NLANR traces, most of the AUCKLAND traces in the figure show some degree of concavity, indicating the existence of a sweet spot at which predictability can be maximized. Generally speaking, these sweet spots are between 8 seconds and 128 seconds for the AUCKLAND traces.

As we noted in the introduction, the existence of the sweet spots here and in the NLANR traces is surprising and contradictory to earlier expectations. Because this effect happens in a large fraction (but not all) of our traces, and the optimal bin sizes vary from trace to trace, we believe that the existence of the sweet spot is not a fluke, but an inherent property of the data.

One question is whether these sweet spots are a property of the AR(32) model. This is not the case. Figure 43 shows the noise-to-signal ratio as a function of bin size for all the predictors on a typical AUCKLAND trace. Each curve corresponds to a predictor, starting from the smallest bin size of 0.125 second to the largest bin size of 1024 seconds. The figure demonstrates that the existence and location of the sweet spot is independent of the choice of predictor.

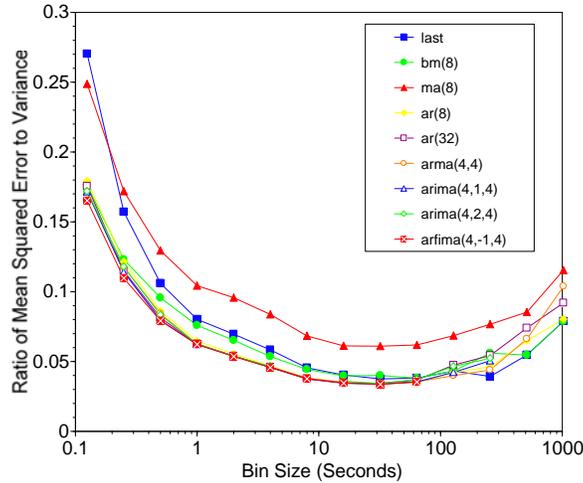


Figure 43: Illustration of “sweet spot” for a typical AUCKLAND trace using different predictors.

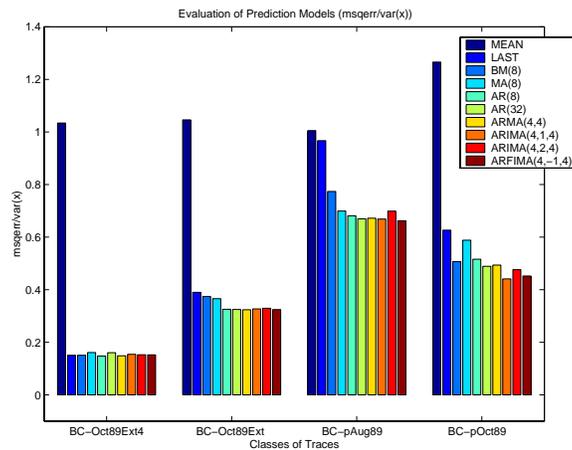


Figure 44: Prediction performance, as measured by the noise-to-signal metric, for 4 BC traces using different predictors. Smaller is better.

5.4 Predicting the BC traces

Our examination of BC traces (Section 4) led us to believe that they would exhibit some predictability. Figure 44 shows the performance, using the of the ten predictors on 4 BC traces, still in terms of the ratio $msqerr/var(x)$. Note that the first two traces are of WAN traffic, while the remaining two are LAN traffic. We used a bin size of 8 seconds for the WAN traces and 0.125 seconds for the LAN traces.

It is clear from the figure that the BC traces do have some predictability, with the two WAN traces being more predictable than the two LAN traces. For the ten predictors, those with an autoregressive component have better performance.

5.5 Traffic prediction example

Before ending our discussion of traffic prediction, let’s take a look at what actual predictions look. As we described earlier, predictors were trained using the first half of the trace data, while predictions are made on the second half. We will use an AR(32) model as our example here. Figure 45 shows traffic prediction on

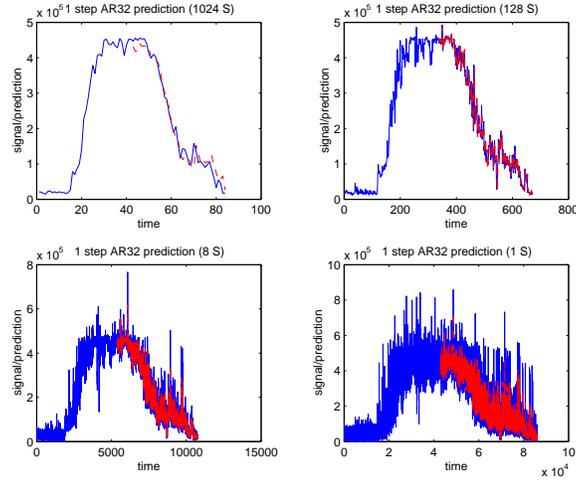


Figure 45: Traffic prediction time-series for a AUCKLAND trace using AR 32 predictor.

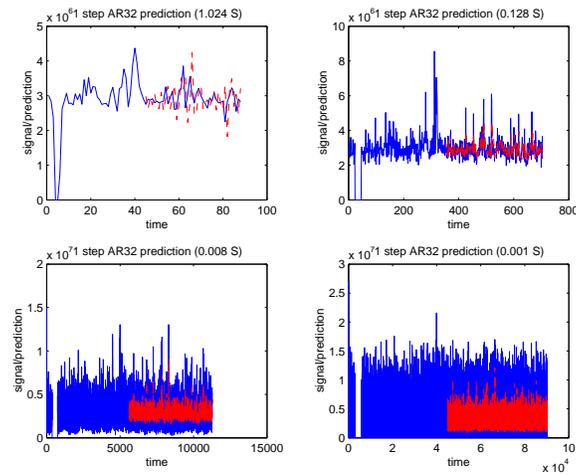


Figure 46: Traffic prediction time-series for a NLANR trace using AR 32 predictor.

an AUCKLAND trace which has good predictability. We show the four bin sizes: 1024 s (upper left), 128 s (upper right), 8 s (lower left), and 1 s (lower right). The blue (or darker) curve is the actual trace. The red (or lighter) curve is the predictions on the second half of the trace. Note that all information up to a point is used to make the prediction of the point. Figure 46 shows a similar family of graphs for an NLANR trace that has some predictability. The bin sizes are: 1.024 s (upper left), 0.128 s (upper right), 0.008 s (lower left), and 0.001 s (lower right).

6 Conclusion

We studied a large number of packet traces (NLANR short-period traces, NLANR long-period (AUCKLAND) traces, and the Bellcore traces), characterizing the consumed bandwidth time series that can be derived from them. We used our characterization to develop a detailed hierarchical classification scheme. We then studied the predictability of representatives of each class. We found the following results:

- Each trace can be easily characterized by a combination of summary statistics, a time-series plot, a

power spectral density (PSD) plot, an autocorrelation function (ACF) plot and a histogram.

- These characteristics can be used to classify traces. These classes can furthermore be arranged hierarchically. We developed a hierarchy that focuses on a trace’s likely predictability using linear models.
- Our classification scheme suggested that about 80% of the NLANR short-period traces have little predictability, while all of AUCKLAND long-period traces have good or some predictability.
- Our evaluation of the predictability of the traces demonstrated that linear prediction models do indeed find the predictability that our classification suggested.
- We found that there was considerable variation in the performance of different predictive models. Generally, models with autoregressive components were best. The performance was also clearly a function of the degree of smoothing of the traffic.
- Predictability does not increase monotonically with smoothing. We found that over 40% of our traces had a “sweet spot”, a degree of smoothing at which predictability was maximized. This is a surprising result.

One issue in our prediction study and characterization is that we are using a methodology that is most appropriate for stationary signals. While many of our traces are clearly stationary, others are clearly not. This is particularly a concern for the prediction study. However, in practice, linear models, including those that don’t model any form of nonstationarity, are likely to be used in online prediction systems. One possible way of quantifying their performance better is to do a randomized evaluation. We also plan to consider the effect of RPS’s managed predictors. Managed predictors are lain on top of the linear predictors we have used here and force a refit of the underlying predictor when prediction error exceeds certain thresholds. In this way, they implement a variant of the family of threshold autoregressive models, which can model many forms of nonstationarity and even non-linearity.

The companion technical report [10] provides considerably more detail on prediction of the traces.

Acknowledgements

We would like to thank Jason Skicewicz and Dong Lu for practical and theoretical help. We would also like to thank NLANR and the Internet Traffic Archive for providing the trace data.

References

- [1] ACM SIGCOMM. Internet Traffic Archive. <http://ita.ee.lbl.gov>.
- [2] BERAN, J. Statistical methods for data with long-range dependence. *Statistical Science* 7, 4 (1992), 404–427.
- [3] BOX, G. E. P., JENKINS, G. M., AND REINSEL, G. *Time Series Analysis: Forecasting and Control*, 3rd ed. Prentice Hall, 1994.
- [4] CAO, J., CLEVELAND, W., LIN, D., AND SUN, D. On the nonstationarity of internet traffic. In *Proceedings of SIGMETRICS 2001* (2001).
- [5] DINDA, P. A., AND O’HALLARON, D. R. An extensible toolkit for resource prediction in distributed systems. Tech. Rep. CMU-CS-99-138, School of Computer Science, Carnegie Mellon University, July 1999.

- [6] GRANGER, C. W. J., AND JOYEUX, R. An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis* 1, 1 (1980), 15–29.
- [7] HOSKING, J. R. M. Fractional differencing. *Biometrika* 68, 1 (1981), 165–176.
- [8] LELAND, W. E., TAQUU, M. S., WILLINGER, W., AND WILSON, D. V. On the self-similar nature of ethernet traffic. In *Proceedings of ACM SIGCOMM '93* (September 1993).
- [9] NATIONAL LABORATORY FOR APPLIED NETWORKING RESEARCH. Nlanr network analysis infrastructure. <http://moat.nlanr.net>. NLANR PMA and AMP datasets are provided by the National Laboratory for Applied Networking Research under NSF Cooperative Agreement ANI-9807579 .
- [10] QIAO, Y., SKICEWICZ, J., AND DINDA, P. Multiscale predictability of network traffic. Tech. Rep. NWU-CS-02-13, Department of Computer Science, Northwestern University, October 2002.
- [11] SANG, A., AND LI, S. Predictability analysis of network traffic. In *Proceedings of INFOCOM 2000* (2000), pp. 342–351.
- [12] WILLINGER, W., TAQUU, M. S., LELAND, W. E., AND WILSON, D. V. Self-similarity in high-speed packet traffic: Analysis and modeling of ethernet traffic measurements. *Statistical Science* 10, 1 (January 1995), 67–85.

Appendix

The following is a list of all of the traces used in this study and the class to which each belongs. The NLANR traces can generally be fetched from website: <http://pma.nlanr.net/Traces/Traces/>, the AUCKLAND long-period traces can be fetched from website: <http://pma.nlanr.net/Traces/long/>, while the BC traces are available at <http://ita.ee.lbl.gov/html/traces.html>.

| Trace Name | Class | Trace Name | Class | Trace Name | Class |
|--------------------|-------|--------------------|-------|--------------------|-------|
| ADV-1017741497-1-1 | 1 | ADV-1017771106-1-1 | 1 | ADV-1017802936-1-1 | 1 |
| ADV-1017869359-1-1 | 1 | ADV-1017880840-1-1 | 1 | ADV-1017901188-1-1 | 1 |
| ADV-1017997629-1-1 | 1 | ADV-1018043715-1-1 | 2d | ADV-1018075534-1-1 | 1 |
| ADV-1018116217-1-1 | 1 | ADV-1018149857-1-1 | 1 | ADV-1018194122-1-1 | 1 |
| ADV-1018235229-1-1 | 1 | ADV-1018299867-1-1 | 1 | AIX-1017771107-1-1 | 2 |
| AIX-1017783988-1-1 | 2f | AIX-1017837934-1-1 | 2f | AIX-1017860079-1-1 | 2 |
| AIX-1017880841-1-1 | 2 | AIX-1017912258-1-1 | 2 | AIX-1017988355-1-1 | 2d |
| AIX-1018032651-1-1 | 2 | AIX-1018075534-1-1 | 2 | AIX-1018086611-1-1 | 2 |
| AIX-1018138787-1-1 | 1 | AIX-1018214469-1-1 | 2 | AIX-1018225954-1-1 | 2 |
| AIX-1018299868-1-1 | 2d | ANL-1017717537-1-1 | 4 | ANL-1017728608-1-1 | 2d |
| ANL-1017815796-1-1 | 2e | ANL-1017837934-1-1 | 4a | ANL-1017880841-1-1 | 2f |
| ANL-1017890102-1-1 | 4a | ANL-1017979097-1-1 | 2 | ANL-1018043716-1-1 | 2d |
| ANL-1018064471-1-1 | 2d | ANL-1018075534-1-1 | 4a | ANL-1018138788-1-1 | 2e |
| ANL-1018181249-1-1 | 2 | ANL-1018225955-1-1 | 3 | ANL-1018290571-1-1 | 2 |
| APN-1017717809-1 | 2f | APN-1017748858-1 | 2e | APN-1017835358-1 | 2a |
| APN-1017848108-1 | 2a | APN-1017943757-1 | 2a | APN-1017956507-1 | 4 |
| APN-1017986458-1 | 4 | APN-1018021107-1 | 4 | APN-1018085707-1 | 4 |
| APN-1018094857-1 | 4 | APN-1018150307-1 | 4 | APN-1018203256-1 | 4 |
| APN-1018224057-1 | 4 | APN-1018302505-1 | 2e | BUF-1017741498-1-1 | 2d |
| BUF-1017750754-1-1 | 2c | BUF-1017826871-1-1 | 2d | BUF-1017869359-1-1 | 2e |
| BUF-1017901187-1-1 | 2e | BUF-1017923309-1-1 | 2d | BUF-1018008706-1-1 | 2d |
| BUF-1018043716-1-1 | 2 | BUF-1018064472-1-1 | 2 | BUF-1018127287-1-1 | 2e |
| BUF-1018194122-1-1 | 2 | BUF-1018214468-1-1 | 2 | BUF-1018235229-1-1 | 2 |
| BUF-1018270250-1-1 | 2 | COS-1017728608-1-1 | 5 | COS-1017760030-1-1 | 2e |
| COS-1017815795-1-1 | 5 | COS-1017860078-1-1 | 5 | COS-1017912257-1-1 | 5 |
| COS-1017943648-1-1 | 5 | COS-1018021588-1-1 | 5 | COS-1018032650-1-1 | 2c |
| COS-1018053396-1-1 | 5 | COS-1018064471-1-1 | 5 | COS-1018194122-1-1 | 5 |
| COS-1018203382-1-1 | 5 | COS-1018235230-1-1 | 5 | COS-1018259188-1-1 | 5 |
| IND-1017750755-1-1 | 2c | IND-1017783988-1-1 | 2c | IND-1017837934-1-1 | 2c |
| IND-1017860078-1-1 | 2c | IND-1017923310-1-1 | 2c | IND-1017932594-1-1 | 2c |
| IND-1017997629-1-1 | 2c | IND-1018043716-1-1 | 2c | IND-1018086611-1-1 | 2c |
| IND-1018116218-1-1 | 2c | IND-1018138788-1-1 | 2c | IND-1018149858-1-1 | 2c |
| IND-1018290571-1-1 | 2c | IND-1018299868-1-1 | 2c | MEM-1017760030-1-1 | 2c |
| MEM-1017783987-1-1 | 2c | MEM-1017826871-1-1 | 2c | MEM-1017869359-1-1 | 2c |
| MEM-1017880840-1-1 | 2c | MEM-1017923310-1-1 | 2c | MEM-1017997629-1-1 | 2c |
| MEM-1018032650-1-1 | 2c | MEM-1018106960-1-1 | 2c | MEM-1018127286-1-1 | 2c |
| MEM-1018138787-1-1 | 2c | MEM-1018181248-1-1 | 2d | MEM-1018246306-1-1 | 2d |
| MEM-1018270251-1-1 | 2 | MRA-1017860081-1-1 | 2b | MRA-1017912260-1-1 | 2b |
| MRA-1017956524-1-1 | 2b | MRA-1018043718-1-1 | 2b | MRA-1018086612-1-1 | 2a |
| MRA-1018106960-1-1 | 2a | MRA-1018149859-1-1 | 2b | MRA-1018259189-1-1 | 2b |
| MRA-1018270252-1-1 | 2a | ODU-1017708659-1 | 2 | ODU-1017748858-1 | 2d |
| ODU-1017795158-1 | 2d | ODU-1017870007-1 | 2d | ODU-1017934607-1 | 2d |
| ODU-1017956506-1 | 2 | ODU-1017977308-1 | 2d | ODU-1017999208-1 | 2d |
| ODU-1018063807-1 | 2 | ODU-1018072957-1 | 2d | ODU-1018172207-1 | 2d |
| ODU-1018203256-1 | 2d | ODU-1018224058-1 | 2d | ODU-1018280607-1 | 2 |
| OSU-1017717808-1 | 6 | OSU-1017770756-1 | 6 | OSU-1017835357-1 | 4 |
| OSU-1017870007-1 | 6 | OSU-1017903557-1 | 6 | OSU-1017934606-1 | 6 |
| OSU-1017977308-1 | 6 | OSU-1018043007-1 | 6 | OSU-1018063807-1 | 6 |
| OSU-1018094857-1 | 2d | OSU-1018181356-1 | 2d | OSU-1018203256-1 | 2 |
| OSU-1018267857-1 | 2d | OSU-1018289755-1 | 2d | TAU-1017717809-1 | 2 |
| TAU-1017730558-1 | 2e | TAU-1017795159-1 | 6 | TAU-1017804308-1 | 2e |
| TAU-1017903557-1 | 2a | TAU-1017934607-1 | 2a | TAU-1018021106-1 | 2a |
| TAU-1018043007-1 | 2a | TAU-1018063807-1 | 2d | TAU-1018107607-1 | 2a |
| TAU-1018141158-1 | 2 | TAU-1018150308-1 | 2f | TAU-1018245957-1 | 2a |
| TAU-1018302507-1 | 2e | TXS-1017708659-1 | 2 | TXS-1017717809-1 | 6 |
| TXS-1017848106-1 | 2f | TXS-1017870007-1 | 2c | TXS-1017890807-1 | 2c |
| TXS-1017934607-1 | 2c | TXS-1017999207-1 | 2c | TXS-1018008356-1 | 2c |
| TXS-1018063807-1 | 2c | TXS-1018116756-1 | 2c | TXS-1018150308-1 | 2c |
| TXS-1018216005-1 | 2c | | | | |

Figure 47: List of all 175 NLANR traces and their classes.

| Trace Name | Class | Trace Name | Class | Trace Name | Class |
|-------------------|-------|-------------------|-------|-------------------|-------|
| 20010220-210122-0 | 3 | 20010220-210122-1 | 2 | 20010221-020000-0 | 1 |
| 20010221-020000-1 | 1e | 20010222-020000-0 | 1 | 20010222-020000-1 | 1a |
| 20010223-020000-0 | 1 | 20010223-020000-1 | 1e | 20010224-020000-0 | 1b |
| 20010224-020000-1 | 2 | 20010225-020000-0 | 1b | 20010225-020000-1 | 1a |
| 20010226-020000-0 | 1 | 20010226-020000-1 | 1 | 20010301-110023-0 | 1d |
| 20010301-110023-1 | 3 | 20010302-020000-0 | 1 | 20010302-020000-1 | 1a |
| 20010303-020000-0 | 1b | 20010303-020000-1 | 2 | 20010304-020000-0 | 1b |
| 20010304-020000-1 | 1a | 20010305-020000-0 | 1 | 20010305-020000-1 | 1a |
| 20010306-020000-0 | 1 | 20010306-020000-1 | 1a | 20010307-020000-0 | 1 |
| 20010307-020000-1 | 1a | 20010308-020000-0 | 1c | 20010308-020000-1 | 1a |
| 20010309-020000-0 | 1 | 20010309-020000-1 | 1e | 20010310-020000-0 | 1b |
| 20010310-020000-1 | 2 | | | | |

Figure 48: List of all 34 AUCKLAND traces and their classes.