Application of Sampling Methodologies to Network Traffic Characterization

Kimberly C. Claffy and George C. Polyzos
kc@cs.ucsd.edu, polyzos@cs.ucsd.edu
Computer Systems Laboratory
University of California, San Diego
La Jolla, CA 92093-0114

Hans-Werner Braun
hwb@sdsc.edu
San Diego Supercomputer Center
San Diego, CA 92186-9784

Abstract

The relative performance of different data collection methods in the assessment of various traffic parameters is significant when the amount of data generated by a complete trace of a traffic interval is computationally overwhelming, and even capturing summary statistics for all traffic is impractical. This paper presents a study of the performance of various methods of sampling in answering questions related to wide area network traffic characterization. Using a packet trace from a network environment that aggregates traffic from a large number of sources, we simulate various sampling approaches, including time-driven and event-driven methods, with both random and deterministic selection patterns, at a variety of granularities. Using several metrics which indicate the similarity between two distributions, we then compare the sampled traces to the parent population. Our results revealed that the time-triggered techniques did not perform as well as the packet-triggered ones. Furthermore, the performance differences within each class (packet-based or time-based techniques) are small.

1 Introduction

Statistics collection in modern networking environments involves cost-benefit tradeoffs. Traditionally, characterizing certain aspects of traffic on wide area networks has been possible by simply maintaining arrays for the distribution of various metrics: packet size, interarrival time, packet type, and geographic flow information. Recent dramatic increases in the speed of wide area backbones pose obstacles to complete statistics collection; managers of high-speed networks are under tremendous pressure to optimize resource usage to fulfill the data collection objective. Sampling offers a strategy to alleviate these pressures.

This research is supported by a grant of the National Science Foundation (NCR-9119473), and a joint study agreement with the International Business Machines, Inc.

Implementing sampling techniques in an operational environment requires a concerted investigation into the effect of sampling on network analysis. This paper presents a detailed study of how accurately various methods of sampling can answer questions related to wide area network traffic characteristics. In Section 2 we present an example of an operational statistics collection mechanism on the current NSFNET that recently resorted to sampling in order to maintain the integrity of its collection. This example illustrates the importance of sampling for networking environments, and provides a rationale for our investigation of sampling methodology.

We then describe an experiment we conducted at an entrance point into the NSFNET backbone environment at which we were able to gather packet traces. These packet traces allowed us to explore the effect of different parameters of sampling, such as: (1) time-driven vs. event-driven methods; (2) random vs. deterministic selection patterns; (3) the granularity, or sampling fraction; (4) the interval, or length of time over which we sample. We use as assessment targets in this study the distribution of packet sizes and packet interarrival times.

2 NSFNET statistics collection

We describe the current implementation of one of the statistics collection processes for the T3 NSFNET backbone, to illustrate an example of a wide area environment faced with data collection demands that have forced the implementation of sampling.

The principal sources of information for the T3 NSFNET backbone come from programs using the Simple Network Management Protocol (SNMP) [4] for simple interface statistics, and specialized software packages for more comprehensive traffic characterization based on traffic type and source/destination. For the T1 backbone, Merit used a modified version of the NNStat [3] package for traffic characterization. Advanced Network Services (ANS) now performs the network operations center (NOC) services for the T3 NSFNET backbone, and designed the

ARTS (ANSnet Router Traffic Statistics) package [2], for traffic characterization. Claffy *et al* [5] [6] provide detailed overviews of statistics collection on the T1 and T3 NSFNET backbones, respectively. We describe here only the specific mechanisms, NNStat on the T1 backbone and ARTS on the T3 backbone, which rely on sampling for traffic characterization.

Each T1 backbone node (NSS) was implemented as a dual token ring interconnecting multiple, typically nine, IBM RT/PC processors. To categorize IP packets entering the backbone based on information contained in packet headers, one RT processor within each NSS was dedicated to examining the header of every packet traversing this intra-NSS processor interconnection facility. This dedicated processor utilized the NNStat package [3] to build statistical objects based on the collected information.

The design of the T3 backbone required significant modification to the statistics collection mechanism. The T3 network design offloaded the packet forwarding process onto intelligent subsystems, consisting of Intel 960 processors with their own memory and firmware. The subsystems can communicate with each other directly via an RS/6000 microchannel bus, allowing them to forward packets without intervention from the main CPU. Because the packet forwarding does not necessarily involve the main processor, accommodating the statistics collection required placing the software which selects IP packets for traffic characterization into the firmware of the subsystems themselves. Each subsystem forwards its selected packets, currently every fiftieth, to the main CPU, where the ARTS software package performs the traffic characterization based on these sampled packets. Note that multiple subsystems, including those connected to T3, Ethernet, and FDDI external interfaces, forward to the RS/6000 processor in parallel.

Although the packet categorization mechanism at each node differs on the two backbones, the backbone-wide centralized collection of the data is the same. Every fifteen minutes, the central agent at the NOC running the collection software queries each of the backbone nodes, which report and then reset their object counters. The collection host is an IBM RS/6000 at the ANS NOC, which during mid-February 1993 was collecting around 25 MB of ARTS traffic characterization data on a typical workday.[1] Table 1 illustrates the traffic characterization objects collected on the T1 and T3 backbones. Note that the T3 backbone only supports collection of the first three objects.

When the collection mechanism maintains sophisticated aggregate objects, even dedicated processors can begin to suffer degradation in the quality of collection under high load. For example, during the early years of the T1 backbone, the utilization was not high enough to strain the

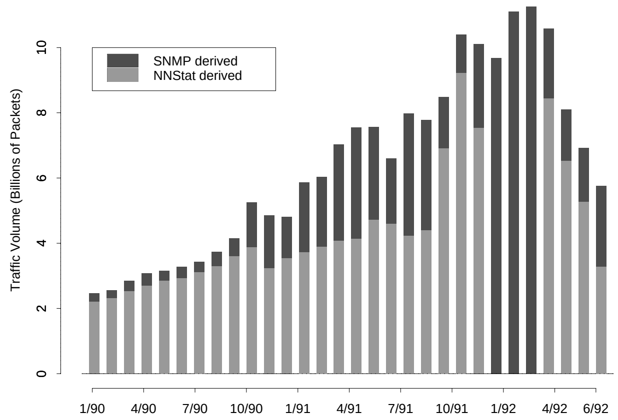

Figure 1: T1 backbone packet totals (billions of packets), as reported independently by SNMP and NNStat, indicate a discrepancy between the two collection processes.

capacity of this dedicated processor. By mid-1991, however, the discrepancies between the SNMP based traffic counts and those derived by means of NNStat had grown to a significant fraction of the total traffic count, as shown in Figure 1. It became clear that the processor collecting the NNStat data was unable to keep up with the total nodal traffic flow.[2]

In September 1991, responding to concerns over the integrity of the data, the operator of the T1 NSFNET backbone deployed a sampling technique which captures only one out of fifty packet headers for traffic characterization purposes. The result was a significant reduction in the discrepancies. Although the sampling imposes a cost of inaccuracies of the traffic signatures, there is no longer complete loss of statistical information during periods of high utilization.

Because each T1 backbone node facility had a processor dedicated to statistics collection, the collection mechanism never imposed a burden on the packet forwarding capacity of the node, although heavy network utilization may have rendered the statistics collecting processor unable to capture all the traffic. In contrast, some components of statistics collection in the T3 architecture are integral to the forwarding process, and therefore may potentially impact the switching capacity. Minimizing overhead in the statistics collection mechanism is essential to the high performance of the T3 backbone.

Although the motivation is different for the T1 and T3 architectures, both statistics collection mechanisms force the consideration of sampling. Future gigabit networks will only intensify the problems. As loads in these envi-

[1] On the T3 backbone, the packet categorization collection mechanism uses a more efficient binary format than that used on the T1 backbone.

[2] Because the SNMP statistics are incremented in the mainstream of packet forwarding, they are more reliable. It is the traffic categorization information, specifically the net matrix, protocol, and port data, which is subject to losses during periods of high utilization.

Table 1: Packet categorization objects on T1 and T3 backbone nodes

Object	**T1**	**T3**
relative to exterior nodal interface		
source-destination traffic volume matrix by network number (packets/bytes)	Y	Y
TCP/UDP port distribution, well-known subset (packets/bytes)	Y	Y
distribution of protocol over IP (e.g., TCP, UDP, ICMP) (packets/bytes)	Y	Y
packet-length histogram at a 50-byte granularity	Y	N/A
packet volume going out of backbone node	Y	N/A
NSS-centric (entire node)		
per second histogram of packet arrival rates (20 pps granularity)	Y	N/A
NSS (intra-NSFNET) transit traffic volume	Y	N/A

ronments outstrip the ability of even dedicated statistics processors to monitor the traffic, sampling will become essential to the integrity of sophisticated data objects which can reflect network usage and behavior.

3 Measurement methodology

We now describe the environment in which we collected the data for our study. The nature of our investigation demands detailed insight into traffic behavior, which requires evaluating each packet traversing the environment. Because NSFNET backbone core nodes typically cannot support the collection of traces capturing all packets over long periods of time, we collected packet traces at a single entrance interface into the backbone. Specifically, we collected a 24-hour trace of packets sent from the SDSC environment to the NSFNET San Diego E-NSS via the FDDI interface. This kind of environment is more conducive to traffic capture than many other points, while still exhibiting a reasonable level of aggregation of traffic. We did not investigate the traffic back from the E-NSS, nor did we investigate the traffic in or out of the E-NSS Ethernet interface.

The 24 hour trace is more than 650MByte long and started at shortly after 22:00PST on the 22 March 1993. Of the 24 hours we created a subset of about one hour, from 13:00 to 14:00 for the 23 March 1993.[3] We then performed sampling simulations on this one-hour trace. Table 2 quantifies the statistics of the per-second packet, byte, and mean packet size distributions for the the data set.

[3]Preliminary experiments for this study used data from the FIX-West interexchange point in Moffet Field, CA. The results of the two data sets were quite similar, but the ENSS data set we use here is more relevant to the current NSFNET statistics collection situation.

4 Sampling mechanisms

Developing a sampling methodology requires an evaluation of the cost and benefit of sampling in the particular domain of study. Our goal is to evaluate the effects of certain sampling parameters on the integrity of the resulting samples. In general, a larger sample can more closely reflect the true parent population, but each instance of sampling imposes a cost, in terms of CPU time, buffer space, and sampling interval, or amount of calendar time one can devote to deriving a particular estimate. The sampling frequency must therefore be weighed against the accuracy requirements and complexity of a given object.

The one-hour packet trace we collected for our experiments represents only a brief interval, indeed itself a sample from the ongoing population of network traffic. For the purposes of our study we treat this packet trace as the true parent population, and the subpopulations drawn by our various sampling techniques as the samples. Standard statistical formulas generally rely on estimates of parameters of the parent population for the default case where the parent population is not known. Because we have access to the actual parameters of this parent population, we use them rather than estimates of them. Our goal is then to assess how close each sample is to its parent population for several key measurements.

Figure 2 illustrates an abstraction of the three main classes of sampling schemes we used in the study: systematic sampling; stratified random sampling; and simple random sampling. For each class, one can implement, or approximate, any particular method via either event-based or timer-based mechanisms. That is, one can use packet counts or timers to trigger the selection of a packet for inclusion in a sample. Implementing these methods at a variety of granularities allows a range of sampling fractions. Furthermore, one can vary the interval over which one samples: for a minute, 15 minutes, an hour, a day, etc. Since the processes are not time-homogeneous, it is not clear that spreading the same number of samples over longer intervals will generate the same result.

Table 2: Summary statistics for distribution of per-second packet and byte volume, and average packet size

Distribution	Min.	25%	Median	75%	Max.	Mean	StdDev.	Skew	Kurtosis
Monday, 22 March 1993 (1.636 million packets during hour)									
Packet arrivals (packets/s)	156	364	412	473	966	424.2	85.1	0.96	4.95
Byte arrivals (kB/s)	26.591	71.1	90.9	117.6	330.6	98.6	38.6	1.2	5.2
Mean per-sec packet size (bytes)	82	190	222	259	398	226.2	50.5	.36	2.9

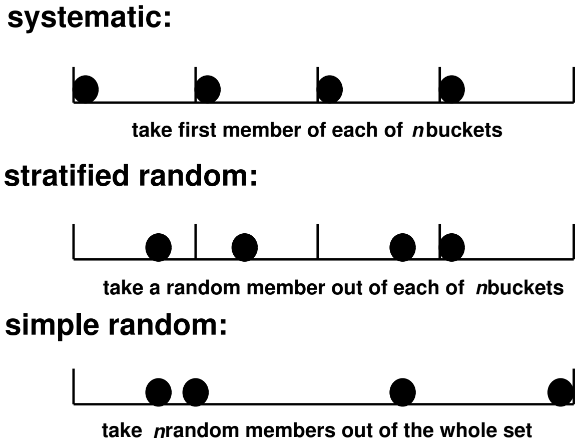

Figure 2: Schematic of Three Sampling Algorithms

We briefly describe each method, first for packet-driven and then timer-driven implementations. The first class of methods, *systematic sampling* involves deterministically selecting every *kth* element (packet) of the data set. *Stratified random sampling* is similar to systematic sampling, except that rather than selecting the first packet from each bucket, a packet is selected randomly from each bucket. Although for both systematic and stratified random sampling the bucket sizes do not necessarily have to be constant, our experiments do use constant sizes. Finally, *simple random sampling* uniformly selects n packets from the total population at random.

Timer-driven sampling methods use a timer rather than a packet counter to trigger the selection of packets to include in the sample. When the timer expires, we select the next packet to arrive. This is a necessary approximation but seemingly inconsequential. As with granularities for packet-based sampling, one selects various time intervals to implement a desired range of sampling fractions.

We implemented all three of the above methods for packet-based sampling, and the first two methods for timer-based sampling, for a total of five basic methods. For each sampling method we selected, we were motivated by an interest in the effects of patterns in the data. We also wanted to determine whether the method currently employed for data collection on the T3 ANSnet backbone, systematically sampling every fiftieth packet, provided significantly different results from simple random sampling. Because preliminary experiments with timer-based sampling were not encouraging, as we explain in Section 7.2 we do not devote much attention to timer methods in this paper.

5 Methodological background

Cochran [7] and Krishnaiah and Rao [11] provide some comparative analyses of which sampling strategies offer lower variance under given conditions. These analyses use the variance of the estimate of the mean as a metric for the sampling method; the lower the expected variance of the estimate, the more *efficient* the sampling method. In our case we are more interested in assessment of the complete distribution. Nonetheless we offer some preliminary insights based on this evaluation mechanism.

If the populations are randomly ordered, we expect all three methods (systematic, stratified, and random) to be equivalent. Systematic sampling spreads the samples more evenly over the population, which can potentially yield greater precision than stratified random sampling. In general, systematic sampling is more precise than simple random sampling if the variance within the systematic samples is larger than the population variance as a whole. If there is positive correlation between pairs of elements within the systematic sample, however, then stratified or simple random sampling will be more efficient.

For populations with a linear trend, stratified random sampling will be more efficient than systematic sampling. Intuitively, one can imagine how if the sample from the first bucket were too low, the sample from each subsequent bucket would also be too low. Stratified random sampling would alleviate this difficulty. Interestingly enough, simple random sampling is less efficient than either systematic or stratified random sampling in this situation [11].

5.1 Theoretical sample size for means

Cochran [7] provides a detailed explanation of the statistical determination of the appropriate random sample size for estimating a given parameter of a population, such as the mean or proportion. We provide an illustration of the appropriate sample sizes to estimate the mean for given confidence levels on the two metrics we selected as analysis targets. As an example we will specify an accuracy of $r = \mp 5\%$ and a confidence level of $100(1-\alpha)\% = 95\%$, which implies z-value of 1.96 in the following formula for the appropriate sample size n:

$$n = (\frac{100z\sigma}{r\mu})^2$$

where μ is the population mean and σ is the population standard deviation.

For our data set (of approximately 1.6 million packets), the packet size distribution had population mean $\mu = 232$ bytes and population standard deviation $\sigma = 236$. These values yield as the appropriate sample size: 1590. Note that these formulas assume sampling from an infinite population, while we are actually using a population of about 1.6 million packets, of which 1,590 constitutes a sampling fraction of around 0.10%. Note that the mean is not a particularly indicative description of the packet size distribution, which is bimodal around 40-byte and 552-byte packets. An accuracy of $r = 1\%$ would require 39,752 samples from the same data set.

For the interarrival time distribution of this data set the population mean is $\mu = 2358$ μsec and the standard deviation is $\sigma = 2734$. These values yield the appropriate sample size: 2066. An accuracy of $r = 1\%$ would require 51,644 samples from this data set.

5.2 Metrics of disparity between distributions

Since both of our characterization targets (and most others) come from distributions for which the mean is not such a helpful description, such estimates are of limited value to us. We seek a more sophisticated assessment of the various metrics, usually obtained through more comprehensive descriptions of the distribution.

Perhaps the best known metric is Pearson's χ^2 statistic, which compares the observed and expected counts within a set of bins which span the range of the data:

$$\chi^2 = \sum_{i=1}^{B} \frac{(O_i - E_i)^2}{E_i}$$

where B is the number of bins, O_i is the number of observations found in the *ith* bin of the sample, and E_i is the number of observations expected in the *ith* bin based on the parent population model. The sampling distribution of χ^2 is approximately the χ^2 distribution where the number of degrees of freedom equals the number of bins minus the number of independent parameters fitted minus one. This approximation improves as the number of counts in each cell increases, and is generally adequate if each cell has at least five expected counts. This statistic is the basis of the χ^2 test, which uses the χ^2 distribution to test hypotheses at specified significance levels about the goodness of fit between a model and a data set.

We performed χ^2 tests for our two target distributions on some of our samples varying several parameters. The results were remarkably compatible with statistical theory. For example, in our experiments for systematically sampling every fiftieth packet, only two or three out of the fifty possible replications produced χ^2 values that would convince a statistician to reject the hypothesis that they were produced by the original distribution at the 0.05 confidence level.

Unfortunately, the χ^2 statistic is sensitive to the size of the data set, making it difficult to compare samples of varying sizes. Therefore, it cannot quantify significant trends when varying the sampling fraction, one of our primary concerns. Goodman and Kruskal [10] note that although useful as a test for the significance of the association between two data sets, the χ^2 statistic, or any simple function of it (e.g., the significance level), cannot serve as a measure of *degree* of association between two sets. On the other hand, we did find significantly higher χ^2 values for the timer-based methods, which motivated us to drop them from the primary focus of our investigation. However, reasonable differentiation among the other methods was not possible with the traditional χ^2 goodness-of-fit testing methods.

Other sophisticated goodness-of-fit tests, such as the Kolmogorov-Smirnov [8] or Anderson-Darling A^2 [1] tests, have proven difficult to apply to wide-area network traffic data [12]. Another disparity metric, which we refer to as *cost*, measures the absolute distance, or l_1 norm, between the expected and observed bin counts: $\sum_{i=1}^{B} |O_i - E_i|$. Consider the following example use of the cost metric. Imagine a network service provider who uses traffic-based charging trying to convince his customers that sampling does not adversely affect their charges. He can offer to reimburse his customers for the difference between their real (if accessible) and observed (i.e., estimated via sampling) traffic. The provider would also like to avoid losing revenue through samples that underestimate the transmitted traffic. If X_i is the number of packets which the network provider attributes to his client based on his sampling, and Y_i is the number of packets which the client actually sent, then there are two possibilities:

- $X_i > Y_i$, in which case client i may express dissatisfaction at being overcharged, or
- $Y_i > X_i$, in which case the service provider loses earned revenue to client i

Note that the actual difference in the number of packets is important here, rather than metrics that compare the general shapes of distributions. Therefore, the provider should use a feasible sampling mechanism that minimizes the l_1 norm. By feasible we assume the comparison of sampling techniques with comparable cost. A service provider might also want a *relative cost* measure, for example the product of l_1 with the sampling fraction, to account for the resource savings of sampling less often.

All these metrics are still subject to the influence of the sample size. Fleiss [9] offers another alternative metric to measure the degree of similarity between two distributions which is free of the influence of the sample size: the ϕ *(*phi*) coefficient*. This metric is derived from the χ^2 metric as follows: $\phi = \sqrt{\frac{\chi^2}{n}}$, where $n = \sum_{i=1}^{B}(E_i + O_i)$. Unlike the χ^2 statistic, which uses the associated χ^2 distribution for hypothesis testing, we are aware of no such corresponding distribution for the ϕ metric.

Paxson [12] considers another χ^2-inspired metric which remains invariant with increasing sample sizes: $X^2 = \sum_{i=1}^{B} \frac{(O_i - E_i)^2}{(E_i)^2}$ and which allows one to compute the "average normalized deviation" across all bins: $k = \sqrt{\frac{X^2}{B}}$.

In the next section we illustrate the application of several of these metrics with an example from our data set, and then select one metric to apply to our data from the domain of interest: a high speed wide area network.

6 Empirical evaluation

For the following example we use a single approximately half-hour (2048 second) interval of packet trace data and sample at exponentially coarser granularities. Figure 3 plots as a function of sampling granularity (inverse of the sampling fraction) the various metrics we described above which indicate the degree of disparity between the sample and the population: the χ^2 metric; the χ^2 significance level (for ease of comparison we plot (1 - the significance level) in the figure); the *cost* and *relative cost (rcost)* metrics; the X^2 metric; and the ϕ metric. Each metric in the figure attempts to measure the goodness of fit of a model to a data set, where in this case our subsamples are the model of the original (in the real world, unknown) data set. According to this figure the *cost*, X^2, and ϕ metrics all exhibit similar behavior. Because the ϕ metric is well established in the statistical literature, we chose this metric for use in our investigation.

We will present results in terms of the range of ϕ-values for a given analysis target, and how these ϕ-values change as we vary one dimension of the parameter space holding the other dimensions constant. A ϕ-value of 0 is consistent with a sample which perfectly reflects the parent population. In general, larger ϕ-values will correspond to poorer samples, i.e., those that diverge more widely from the sampled population. When a network operator selects a sampling method, with an associated sampling fraction and interval, he buys a certain range of ϕ-values which will characterize his samples. Although we do not offer a precise threshold below which all ϕ-values are acceptable, we do offer suggestions for how the ϕ-value scale can guide a sampling methodology.

A complication in our experiment is the fact that some of the samples share members with other samples, and thus there is correlation among the samples. This correlation inhibits statistically precise statements about the superiority of one sampling method over another. On the other hand this approach does allow us to easily order sampling methods based on their performance.

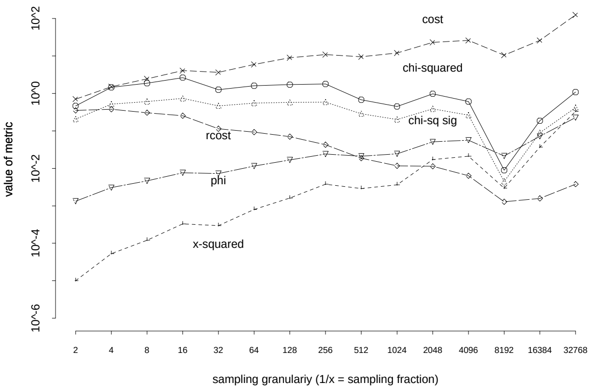

Figure 3: Various metrics of disparity for samples as a function of exponentially increasing sampling granularities

7 Application of methodology

Now that we have presented our methodology for scoring the samples for each target, we concentrate on the effects of the various sampling parameters in isolation. Our experiment consists of a large number of samples exploring the domain based on:

1. class of sampling method (systematic, stratified random, simple random)
2. time-driven vs. event-driven methods
3. granularity, or sampling fraction
4. the interval, or length of time over which we sample

The first two dimensions cover the range of sampling methods which we employ. The latter two dimensions allow further subdivisions to the parameter space. We ran five replications for each method to avoid misleading outlying samples.

We apply our evaluation methodology to the analysis of two distributions: packet size and interarrival times. We show samples which reflect the true population to varying degrees. We then provide graphs which show the effect of varying a single parameter on the range of scores. The objective is to provide a framework for evaluating what count as good or bad ϕ-value scores, and to demonstrate our analysis for the selected targets.

In our sampling simulations we use an exponentially increasing time window relative to the beginning of the

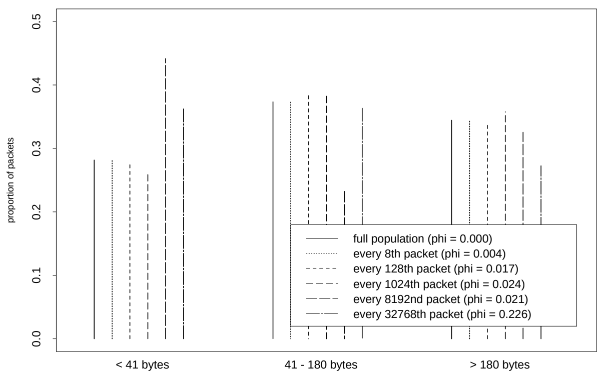

Figure 4: Distribution of packet sizes as a function of five sampling granularities (1024 second interval, systematic sampling)

hour-long trace. To modulate both the time windows as well as the sampling interval, we also ran samples at exponentially decreasing sampling fractions, starting at every other packet, and decreasing the fraction down to one in 32,768 packets. We then binned the interarrival time and packet length distributions for use in our χ^2 based statistic calculation, as we describe below.

7.1 Bin selection

Calculation of the χ^2-based metric that scores our individual samples requires the selection of bins, or ranges, in which to group the data sets. In this section we present the ranges that we used for our two targets, and histograms which illustrate the distributions over these ranges. Table 3 provides summary statistics for the full population for both the packet size and interarrival time distributions.

7.1.1 Packet size distribution

To compare the packet size distributions, i.e., the number of bytes per packet, we compared the proportion of packets within the following three ranges (in bytes): less than 41; between 41 and 180; and greater than 180. We chose these bins based on our knowledge of the typical packet size distribution of network traffic. We experimented with bin sizes which accounted for a fairly large number of packets, and also which characterize certain protocols: ACKs, character echos, transaction-oriented, bulk transfer. Figure 4 compares the distribution of packet sizes into these bins at five sampling granularities.

7.1.2 Interarrival time distribution

For the packet interarrival time distribution, we used the following bins (in μsec): less than 800 μs; between 800 and 1199 μs; between 1200 and 2399 μs; between 2400 and 3599 μs; and greater than 3600 μs. We chose these bins to achieve an relatively even distribution of data among them. Figure 5 shows a histogram of several samples of packet interarrival times dividing them into these ranges. The increasing ϕ-value scores shown in the legend reflect the divergence in the sample accuracy as the sampling fraction decreases. We discuss these scores in detail in the next section. Table 3 summarizes the parameters of the full hour packet population, subject to the 400 microsecond clock granularity described in Section 3.

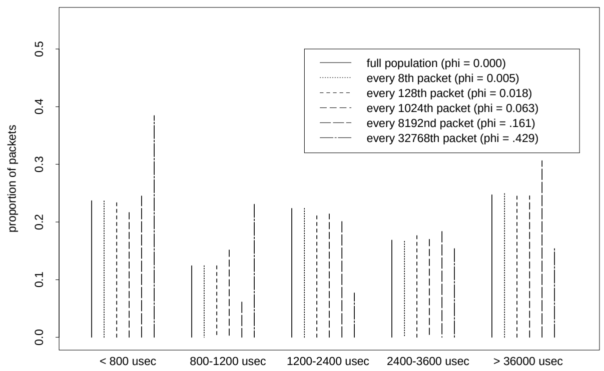

Figure 5: Distribution of packet interarrival times as a function of five systematic sampling granularities (1024 second interval)

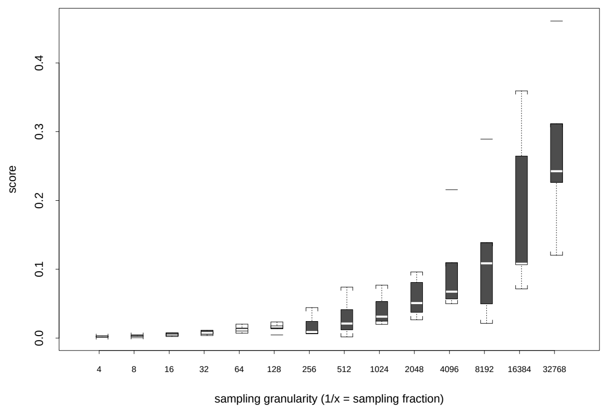

Figure 6: Ranges of systematic sampling ϕ-value scores for packet size distribution as a function of sampling fraction for 1024 second interval (1/x = bucket size)

7.2 Sampling fraction and method

Using these bins to base our scoring, we investigated the variation of individual sampling parameters. To examine the effect of the sampling fraction, we first focused on one method, systematic sampling, and ran several replications of this method at a range of sampling fractions. To achieve a wider range of replications for systematic sam-

Table 3: Summary statistics for distribution of packet sizes and interarrival times

Min.	5%	25%	Median	75%	95%	Max.	Mean	Std.Dev.
Total Population = 1.63 million packets								
packet size								
28	40	40	76	552	552	1500	232	236
packet interarrival times								
< 400	< 400	400	1600	3200	7600	49600	2358	2734

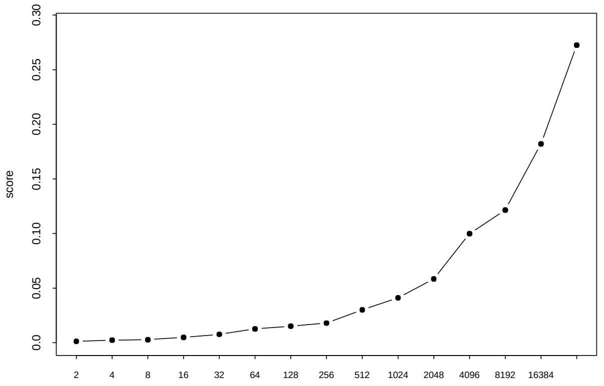

Figure 7: Means of systematic sampling ϕ-value scores for packet size as a function of sampling fraction for 1024 second interval (1/x = bucket size)

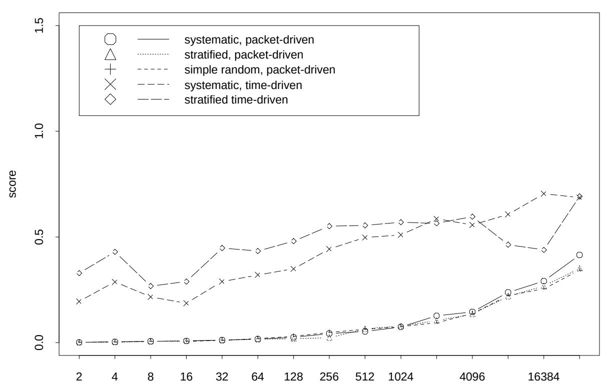

Figure 9: Mean sample ϕ-value scores as a function of sampling fraction for packet interarrival time distribution

ples, we varied the point within the data set at which to begin the sampling procedure. The boxplots[4] in Figure 6 show the range of ϕ-value scores for each systematic sample for the packet size distribution assessment. The x-axis corresponds to the sampling granularity, or the reciprocal of the sampling fraction. The first box plot on the left corresponds to every fourth packet, and most of the scores are near perfect zeros. The figure shows two clear effects of decreasing the sampling fraction, and holds with other methods as well: increasing values, which indicate poorer snapshots of the parent population; and increasing variance within the set of samples for each method. Figure 7 shows the means of the boxplots in Figure 6.

To illustrate the effect of the sampling method, we used the five methods in our experiment to assess the packet size distribution. Like the boxplots in Figure 6, Figure 8 indicates the effect of increasing the sampling fraction on the ability of the sample to estimate the true population.

Figure 9 illustrates the same metric for the packet interarrival time distribution. Figures 8 and 9 show for two targets a general trend which we expect holds for other targets as well: there is little difference in performance among the packet-based methods, and the timer-based methods are uniformly worse. Timer-based sampling is particularly bad for assessing interarrival times, since one tends to miss bursty periods with many packets of relatively small interarrival times, and thus tends to skew the true interarrival distribution toward the larger values. For the remainder of the discussion, we thus restricted ourselves to only two of the packet based methods: systematic and stratified random sampling.

[4]In a boxplot, the dotted lines (or "whiskers") from the bottom to the top of the box, extend to the extreme values of data or 1.5 times the interquartile difference from the center, whichever is less.

7.3 Length of interval

We have investigated how the sampling fraction affects the sample size; another way to increase the sample size, and thus allow greater accuracy in any desired estimate of the parent population, is the duration of the sampling interval. However, network traffic is typically non-stationary, and so the effect of spreading the sampled packets over a longer interval is not clear.

In order to experiment with the effect of the interval, we chose one particular sampling method, systematic sampling, and varied the interval during different runs of that method. Figures 10 and 11 show the resulting ϕ-value scores for the packet size and interarrival time distributions for our data set. Although the left side of these figures reflect smaller time intervals and are noisier, one can see general trends at later intervals. For all sampling fractions the sampling scores improve with elapsed time, as one might expect.

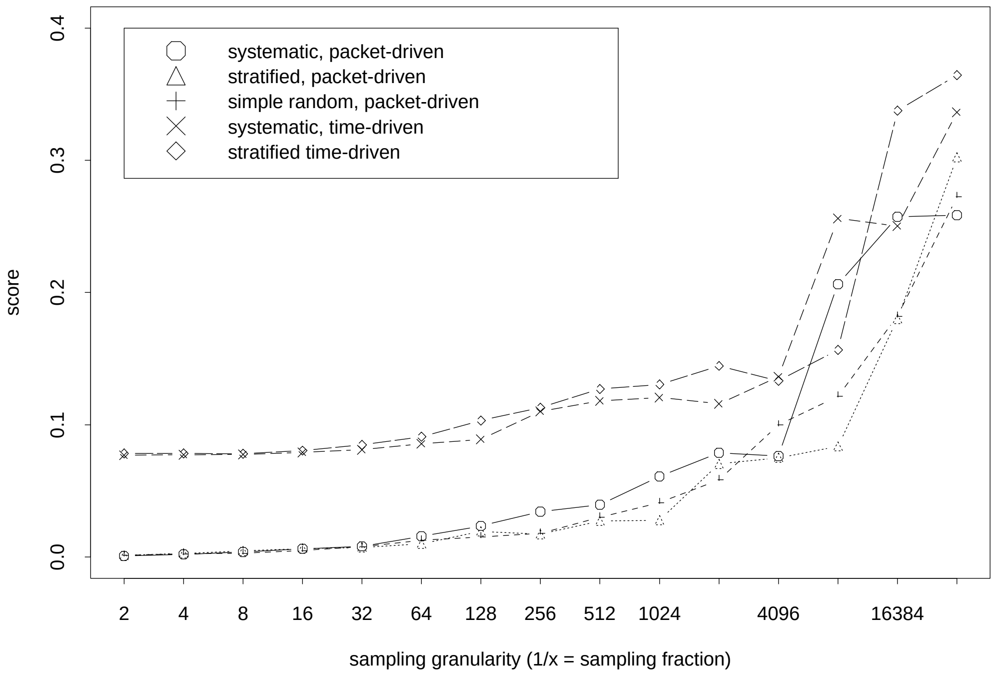

Figure 8: Mean sample ϕ-value scores as a function of sampling fraction for packet size distribution

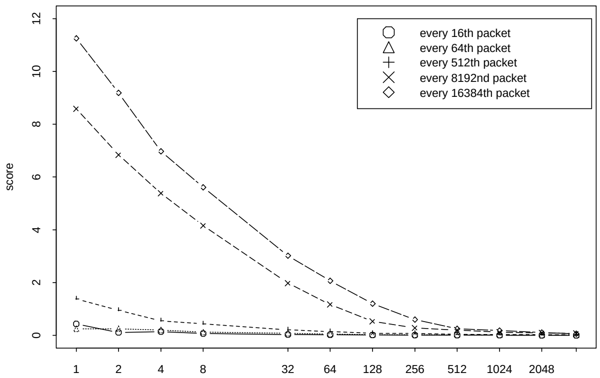

Figure 10: Mean systematic sample ϕ-value scores for packet size distribution as a function of elapsed time (in minutes)

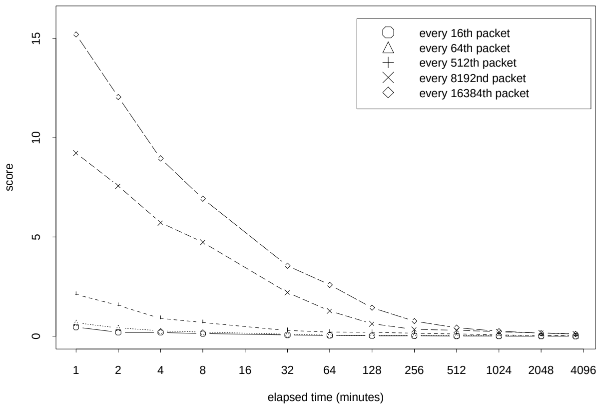

Figure 11: Mean systematic sample ϕ-value scores for packet interarrival time distribution as a function of elapsed time (in minutes)

8 Conclusions

We have presented a framework for the empirical evaluation of sampling techniques for network traffic characterization. We have then applied our methodology to two target metrics: distribution of packet sizes, and distribution of packet interarrival times.

Our experimental data consisted of a packet trace obtained from an entrance interface into the NSFNET national backbone. Because the characteristics of our populations of network data do not fit into any categories analyzed in the literature, we offer in this paper empirical data on sampling simulations run on an isolated packet trace while controlling various experimental parameters.

We have applied the traditional χ^2 test to evaluate the goodness of fit of the sampled distribution to the original complete distribution. One important result is that the current technique of systematic sampling used for statistics collection on the NSFNET backbone provides samples that are compatible with the original distribution of packet sizes and interarrival times at the 0.05 significance level.

Because the χ^2 technique is sensitive to the sample size, and therefore inappropriate for comparison of samples of different sizes, we have focused our evaluation on the ϕ metric which measures similar deviation but is not sensitive to the size of the sample. Based on this metric, we have considered systematic, stratified random, and random sampling by packet or time and various sampling fractions and sampling intervals.

Our results revealed that the time-triggered techniques did not perform as well as the packet-triggered ones. Furthermore, the performance differences within each class (packet-based or time-based techniques) are small. The ϕ^2 metric characterizes the degree of association between the sample distributions and the population, but it does not provide absolute characterizations of sampling performance and in particular it is not conducive to rigorous hypothesis testing. However, it is a useful tool to demonstrate that a technique is generally superior to another across sampling fractions and sampling intervals.

Our methodology can be extended and applied to characterizations of network traffic that are based on proportions, e.g., TCP/UDP port distribution. More difficult would be to characterize the goodness of fit of the sampled source-destination traffic matrix, mainly because of its large size and because many traffic pairs generate small amounts of traffic during typical sampling intervals.

9 Acknowledgements

We would like to express our appreciation for the cooperation and assistance with data collection we received for earlier versions of this paper from Milo Medin of NASA Ames Research Center. We also would like to acknowledge the helpful discussions and exchange of ideas on this topic with Bilal Chinoy of SDSC, Jordan Becker and David Bolen of ANS, and Mark Knopper and Susan Horvath of Merit, Inc.

References

[1] T. W. Anderson and D.A. Darling. Asymptotic theory of certain goodness-of-fit criteria based on stochastic processes. *Annals of Mathematical Statistics*, 23:193–212, 1954.

[2] ANS. ARTS: ANSnet Router Statistics software, 1992.

[3] R.T. Braden and A. DeSchon. NNStat: Internet statistics collection package. Introduction and User Guide. Technical Report RR-88-206, ISI, USC, 1988. Available for a-ftp from `isi.edu`.

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

[5] K. Claffy, H.-W. Braun, and G. C. Polyzos. Tracking long-term growth of the nsfnet backbone. In *Proc. INET '93, San Francisco, CA*, August 1993.

[6] K. Claffy, G. C. Polyzos, and H.-W. Braun. Traffic characteristics of the T1 NSFNET backbone. In *Proc. INFOCOM '93, San Francisco, CA*, April 1993.

[7] W. Cochran. *Sampling Techniques.* John Wiley & Sons, 1987.

[8] R. B. D'Agostino and M. A. Stevens, editors. *Goodness of Fit.* Marcel Dekker, Inc., 1986.

[9] J. Fleiss. *Statistical Methods for Rates and Proportions.* John Wiley & Sons, 1981.

[10] L. Goodman and W. Kruskal. Measures of association for cross classifications. *Journal of the American Statistical Association*, pages 732–763, December 1954.

[11] P.R. Krishnaiah and C.R. Rao. *Handbook of Statistics, Volume 6: Sampling.* North-Holland, 1988.

[12] V. Paxson. Empirically-Derived Analytic Models of Wide Area TCP Connections. Master's thesis, UC, Berkeley and Lawrence Livermore National Laboratory, 1992.