Traffic Behavior Analysis with Poisson Sampling on High-speed Network[1]

Guang Cheng，Jian Gong
(Computer Department of Southeast University Nanjing 210096, P.R.China)

Abstract: With the subsequent increasing of backbone bandwidth, the full trace capture will be seriously hindered. The research in the paper results from a study analyzing packet traces obtained from a backbone link in CERNET. Millions of packets are passed per minute through the link that makes the traffic measurement difficult or even impossible. Poisson sampling measure proposed by RFC2330 is used in the research to alleviate the measurement suffering while preserving the statistically representative characteristics of the traffic population. Algorithms for the continuous process sampling and the discrete process sampling are proposed in the paper. The relationship between the mean packet length of the Poisson sampling data sets and the test data set is discussed to show the advantage of the algorithms. The network traffic throughput can be estimated according to the mean length of sampling packets, Poisson rate and during sampling time.

Keywords: Poisson Sampling, Traffic Analysis, Network Measurement, Traffic Throughput

1. Introduction

Statistical collection in modern networking environment 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, packet type, etc.. As the Internet is rapidly growing in the number of users, traffic levels, and topological complexity, these developments render the characterization of network usage and workloads more difficult, and yet more critical. The dramatic increases in the Internet bandwidth make the statistical collection more difficult or even impossible, and mangers of high-speed networks are under tremendous pressure to optimize resource usage to fulfill the data collection task. Sampling offers a strategy to alleviate these sufferings [6].

In the paper, a sampling methodology for the analysis network traffic traces collected from a backbone link in CERNET is developed. Analyzing the network traffic traces in such an environment is problematic due to the vast amount of data available. In every minute, millions of packets pass through the link, so that most of the software packages have problems handling even the simplest of analysis task, such as sorting. One solution, also a challenge, is to select a sample from the traffic that will effectively preserve the statistically representative characteristics of the whole traffic population. Previous studies on the statistics of the traffic did not use a specific strategy to select their samples. However, there is a need for robust sampling strategy that can be applied to analyze large traffic traces.

The paper discusses a random sampling strategy based on Poisson sampling proposed by RFC2330, to tackle the task of choosing a sample of manageable size and statistically representative of the full traffic. Poisson sampling is appropriate due to the properties it has. The most important property of Poisson sampling is that the query arrival process should not necessarily conform to the Poisson process that enables unbiased sampling for any stochastic arrival process [7]. An effective sampling strategy will enable the analysis of bigger data sets. Moreover, a Poisson based sample selection methodology will increase the probability that a sample would represent the entire data set in preserving its statistical characteristics.

Algorithms for the continuous process sampling and the discrete process sampling are given in the paper, which are the core of the Poisson sampling implementation. The relationship between the mean packet length of the Poisson sampling data sets and the test data set is discussed to show the advantage of the algorithms. The network traffic throughput can be estimated according to the mean length of sampling packets, the Poisson rate and the sample time.

The paper is organized as follows. Section 2 describes collecting CERNET traffic method and section 3 reviews network traffic sampling methods. Section 4 describes Poisson sampling methodology and proposed the algorithms of the continuous process

[1] This research was supported by 863 project of china under grant No. 863-317-01-33-99.

sampling and the discrete process sampling. Section 5 describes the traffic performance analysis of Poisson traffic samples. Section 6 is a conclusion followed by references.

2. Traffic Collection

The raw traffic data used in the paper comes from CERNET Eastern China (north) regional network that covers three provinces and connects about 150 universities and schools. When the connection between CERNET backbone and the regional network was upgraded to OC-48, the network traffic level moved up from 12000 pps to 35000 pps rapidly and the ratio of the zenith of traffic level and the lowest point of traffic level ascended 4 from 1.5. To monitor the behavior of a link in such a high speed, the famous Libcap packet capture library [3] was used.

At present, the traffic measurement system is composed of two hosts, the Traffic Capture Host and the Traffic Classification Host. The Traffic Capture host captures all packets that pass through a backbone link of CERNET, intercepts the packet headers, and sends these encapsulated packet headers to the Traffic Classification Host using UDP. The Traffic Classification Host reads the header information of each header and classifies it into its traffic flow. Netflow measurement is not chosen because this measurement is transparent to the backbone router, so that the performance of the router is not effected.

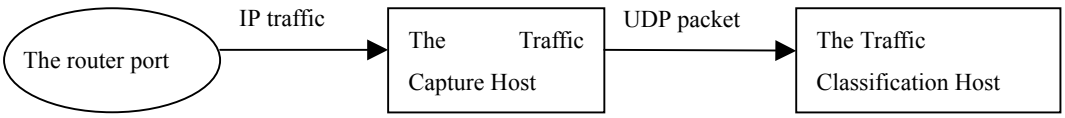

Figure 1: The Measurement Architecture of high-speed IP network Traffic

3. Review of network traffic Sampling Methods

Statistical sampling of network traffic was first used by Claffy et al. [1] in the early 1990's for traffic measurement on the NSFNET backbone. Claffy studied classical event and time driven sampling methods to reduce the number of packets that would need to be received and processed by a network management node. Sampling every 50^{th} packet resulted in a significant improvement in accuracy of statistics taken under high utilization periods [1, 5]. Claffy describe the application of event and timed-based sampling as network traffic measurement. The three algorithms studied were, systematic, stratified random, and simple random [1]. In event based or "packet triggered" sampling, packets are counted to determine when the next packet is to be sampled, whereas in timer-based sampling a "timer trigger" is used. Time-based sampling was shown to be less accurate than event-based. In RFC2330 [4], there are three different methods for approximating Poisson timer-based sampling. The three methods have their own problems so in the study, we use event-based sampling methods. Of the event-based sampling methods, all three algorithms could be used for characterizing network traffic.

4. Poisson Sampling Methodology

Sampling with fixed interval is the simplest scheme. The disadvantage of fixed interval sampling is the elimination of selection chance for some instances of data points. The bias created by fixed interval sampling increases significantly if the sampling interval is increased. RFC 2330 [4] discuss two potential problems with fixed sampling. If the statistics being measured itself exhibits periodic behavior then there is a possibility that the sampling will observe only part of the periodic behavior if the periods happen to agree. The other problem is that the act of measurement can perturb what is being measured. Therefore, a random sampling strategy is required for assigning a probability for selection. However, the most critical issue is to determine the random stochastic process to use for random selection. The Poisson sampling process is the most suitable random sampling process as it includes the following properties.

Firstly, it is unbiased. When Poisson sampling is applied, all instances of a stochastic arrival process would have equal chance of selection [7]. Secondly, it is proportional. The characteristics of the arrival process for the traffic packets may change due to some

factors such as time of day, day of the week, etc. These factors may stay in effect for different time periods during the sampling process. Changes in the effective parameters create different stages of the stochastic process, such as morning stages, Spring Festival stages, etc. The advantage of the Poisson sampling process is that the duration length ratio of stages is captured in the ratio of number of observations taken during the stages, allowing proportional sampling of the different stages of the observed stochastic process [8]. Thirdly, the heterogeneous Poisson sampling arrivals are comparable. If $\{N_1(t), t \geqslant 0\}$ and $\{N_2(t), t \geqslant 0\}$ are both Poisson processes with rates λ_1 and λ_2 respectively ($\lambda_1 \neq \lambda_2$), then the averages of the samples obtained by $N_1(t)$ and $N_2(t)$ can be compared or combined depending on the time interval they have been applied [8]. This property allows utilization of different studies that apply Poisson sampling even if they use different parameters. Finally, there is flexibility on the stochastic arrival process from which the sample is selected. Poisson sampling can be applied to any kind of stochastic arrival process [8]. This is the most important property for applicability of Poisson sampling on the traffic traces, since there is no information on the type of stochastic arrival process of the traffic packets.

Poisson sampling can be applied in two different cases: continuous process sampling and discrete process sampling. For continuous process sampling, selection of the next sample point is comparatively easy. The random number of the next sample is generated according to an exponential distribution with parameter λ (inter-arrival number of the next sample $x \backsim \mathrm{Exp}(\lambda)$). The formulation of the random number generator for exponential distribution can be derived from the cumulative density function (cdf) of the exponential distribution, given in Equation 1.

$$F(x) = \int_{-\infty}^{x} f(y)dy = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases} \qquad (1)$$

If x is generated according to an exponential distribution, then the outcome of cdf, F(x) for $x \geqslant 0$, has a Uniform(0, 1) distribution. By calculating the analytical inverse of the exponential cdf in Equation 1, we can develop the desired formula, which is stated in Equation 2. After each sample point, a new uniform number u has to be generated to calculate the next exponentially distributed inter-arrival time using Equation 2.

$$F^{-1}(u) = \begin{cases} -(1/\lambda) * \ln(1-u), & 0 \leq u \leq 1 \\ 0, & u < 0 \text{ or } u > 1 \end{cases} \qquad (2)$$

The other case of the Poisson sampling, discrete process sampling is used where the stochastic process under observations has discrete arrivals. For discrete stochastic arrival traffic packets, sampling is done by randomly generating a number $u \backsim \mathrm{Uniform}(0, 1)$ and then find the corresponding n, the number of arrivals to skip before the next sample, using Poisson process with parameter $\lambda > 0$, $\{N_1(t), t \geqslant 0\}$. The probability mass function of the Poisson process is given in Equation 3.

$$F(x) = \frac{\lambda^k \exp(-\lambda)}{k!}, \qquad \lambda > 0,\ k = 0, 1, \ldots, \qquad (3)$$

However, the analytical inverse of the Equation 3 is not available. Therefore the following algorithm is used to generate the Poisson variant n [7].

First step: set j = 0 and $y_j = u_0$, where $u_j \backsim \mathrm{Uniform}(0, 1)$, j = 0, 1, … ,

Second step: if $y_j \leqslant \exp(-\lambda)$, return n = j and stop.

Last step: j = j+1, and $y_j = u_j y_{j-1}$, Goto second step.

As in the continuous sampling case, another random n is generated using the algorithm stated above.

5. Sampling traffic performance analysis

5.1 sampling traffic results

In order to illuminate the algorithm of Poisson sampling, now we study the relation between the mean length of packets of the Poisson sampling data sets and of the test data set. The entire traffic population consists of all traffic packets that pass through the monitored link all the time. For the analysis simplification, 10,000 continuous traffic packets are chosen as the test data set, and the selected samples will be

Table 1: 10 sample set of Poisson sampling

Sample Set	λ	n	$\bar{X}$	S
1	1	5057	778	687.3
2	2	3394	779	684.7
3	4	1982	783	687.9
4	8	1124	798	692.5
5	16	587	803	679.2
6	32	303	769	702.3
7	64	149	770	691.8
8	100	98	722	681.2
9	120	95	738	688.5
10	140	94	722	681.3

used to estimate the characteristics of this data set. In the traffic packets traces structure, the entries are given in the order they arrive.

Ten different sample sets were generated from the test data set using different sampling rates for Poisson sampling. The sampling rate is equivalent to the mean used for Poisson sampling, hence is the number of arrivals between two sample points. As the mean used for Poisson sampling increases, the number of observations skipped between the data units that contribute to the sample set increases either, implying a decrease in sample size. Table 1 shows the sampling rate applied for Poisson sampling and the sample size for each individual sample. Therefore, each individual sample set is compared with the entire test data set. The mean length of packets obtained from the data set is 781.8 with a standard deviation of 687.1. The mean length of packets obtained from each sample and the standard deviation of the sample of traffic packets are listed in Table 1. Suppose the traffic packet population mean packet length is μ_0, and its standard deviation σ_0, they can be shown as Equation 4.

$$\mu_0 = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \sigma_0 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_0)^2 \tag{4}$$

In Equation 4, x_i is the length of the packet, and N is the size of the traffic packet population. We also can get the mean packet length $\bar{X}$ of the sample set and its standard deviation S according to the Equation 5 that is the unbiased parameter estimation of the traffic population.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{5}$$

In Equation 5, n is the size of the traffic packet sample set.

From the Table 1, we can know that when λ <100, the number of the sample set doesn't decrease continuously. Because the function rand() of linux return a pseudo-random integer between 0 and RAND_MAX, based on the problem, the maximum λ value that we will choose for the sampling study, should be less than 100.

5.2 Mean Length of Packets Hypothesis Tests

When obtained the estimated value of parameter, hypothesis testing at the 5% and 1% significance levels must be used to test the statistical significance of the difference in mean length of packets between each sample set and the entire data set. The hypotheses are as follows Equation 6,

$$\begin{aligned} H_0 &= \mu_i = \mu \\ H_1 &= \mu_i \neq \mu \end{aligned} \tag{6}$$

where μ is the mean length of packets obtained from the data set and μ_i is the mean length of packets obtained from sample set i, i = 1, …, 10. Each sample is tested with respect to the entire data set, considering the mean length from the data set as a constant. Hence, the tests are based on comparing a sample mean to a specific value. Because the standard deviation σ_0 is known and the sample size is larger than 30, so we choose N(0, 1) as the test statistics. The test statistics is Equation 7.

$$U = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1) \tag{7}$$

Given 5% and 1% significance levels, we can get the $u_{0.025}$ = 1.95 and $u_{0.005}$ = 2.575 according to additive Table A in [8]. Table 2 displays the test statistics testing the statistical significance of the difference between the mean length of packets from each sample and the mean length of packets from the entire data set. The critical values for the relevant tests and the results of the hypothesis testing on the 5% and 1% significance levels. It can be observed that the difference between the mean of packets of each sample set and the test data set are statistically insignificant. In the other words, sample size larger than 100 can be used in traffic statistical study.

Table 2: the hypothesis tests of mean length of packets

Sampling set	Test Value	Results of hypothesis testing	
		95%	99%
1	-0.393	cannot reject H_0	Cannot reject H_0
2	-0.237	cannot reject H_0	Cannot reject H_0
3	0.078	cannot reject H_0	Cannot reject H_0
4	0.790	cannot reject H_0	Cannot reject H_0
5	0.748	cannot reject H_0	Cannot reject H_0
6	-0.324	cannot reject H_0	Cannot reject H_0
7	-0.210	cannot reject H_0	Cannot reject H_0
8	-0.862	cannot reject H_0	Cannot reject H_0
9	-0.621	cannot reject H_0	Cannot reject H_0
10	-0.844	cannot reject H_0	Cannot reject H_0

5.3 Theoretical Sample Size for Mean Length

According to the precision, now we consider the choice of the sample size. The statistical determination of the appropriate random sample size for estimating the mean parameter of a population is provided in [9]. The appropriate sample size n can be calculated as Equation 8,

$$n \geq \frac{N\sigma^2}{N(\varepsilon \times \mu)^2 + \sigma^2} \quad (8)$$

where N is the population size, σ is the population standard deviation, ε is the specified accuracy, μ is the population mean.

For our test data set, the packet size distribution had population mean μ =782 bytes, population standard deviation σ =687.1, and population size N=10,000. With accuracy of ε = 5%, according to Equation 8, the appropriate sample size is larger than 300, so the sampling rate for Poisson Sampling is about 32. If we choose ε = 1%, then the sample size must large than 4357. The population size of the Equation 8 is limited. But for population size that is unlimited, we can induce the Equation 9 that can deal with the unlimited population size on the Equation 8.

$$n \geq \frac{\sigma^2}{(\varepsilon \times \mu)^2} \quad (9)$$

The sign of the Equation 9 is the same as the Equation 8. And such as the above, we hypothesis that μ =782 bytes, and σ =687.1, and ε = 5%, then the sample size must be larger than 309 if the population size is unlimited.

5.4 The Estimation of Traffic Throughput

According to the discussion of the above paragraphs, the estimated traffic throughput can be computed by the Equation 10 in the range of the precision of sampling parameters.

$$throughput = \frac{\mu \times n \times (\lambda + 1) \times 8}{during} \quad (10)$$

where μ is the mean length of sampling packets, n is the sampling size, λ is the Poisson sample rate, and during is the during time of measuring traffic sample. For example, in the paper, if the chosen precision ε =0.05, then μ =769 bytes, n= 303, λ =32, and during = 0.233s, so the traffic throughput is 264MB/S during the measuring traffic according to the Equation 10.

5 Conclusion

Successful modeling of network traffic behavior can be achieved by effective analysis of the traffic data collected from the backbone router. Millions of collected data creates enormous traffic packet traces, the sizes of which are increasing exponentially as the network traffic monitored over the time, so that with current computational tools, even the simplest analytical tasks are difficult to perform. To solve the problem, it is reasonable to develop a sampling methodology for analyzing large traffic data sets. The challenge is how to choose samples effectively from the traffic population, which carry the statistical characteristics of the entire traffic data. Such a sample will reduce the size of the data to be collected and handled by network manager. This paper proposes a random sampling algorithm based on Poisson sampling that can tackle the task of choosing a sample set of manageable size to represent the whole data set. Poisson sampling is chosen because it does not require the traffic packet arrivals to conform to a particular stochastic arrival process and provides unbiased sampling.

A test data set is collected from the real traffic from the CERNET backbone. Ten different samples are chosen from the entire data set, using difference Poisson sampling rate, and the statistical significance of the difference between the mean length of packets from the 10 samples and the test data set are tested. The tests have shown that the difference between the means of packets from each sample set and the test data set were not statistically significant. we find that when λ <100, the number of the sample set doesn't decrease continuously.

Poisson sampling is a successful sampling strategy that will create the opportunity to analyze large traffic traces using significantly reduced sample sizes without losing the statistical characteristics of the traffic population. More detailed analysis of data can be performed with less computational effort, which can provide the opportunity to apply types of analysis that are impossible due to today's software and hardware capacity limitations. Better and more detailed analysis of traffic traces will lead to better understanding of the high-speed network traffic behavior, and result in managed and devised network services.

References

[1] K. Claffy, G. Polyzos, and H. Braum, "Application of sampling Methodologies to Network Traffic Characterization", Computer Communication Review, Vol. 23, No. 4, pp. 194-203,1993.

[2] Kevin Thompson, gregory J. Miller, and Rick wilder, "Wide-Area Internet Traffic Patterns and characteristics (Extended Version)", IEEE Network, November/December 1997.

[3] pcap - Packet Capture library, ftp://ftp.ee.lbl.gov/libpcap.tar.Z, June 1998.

[4] V. Paxson, G. Almes, J. Mahdavi, M. Mathis, Framework for IP Performance Metrics, IETF RFC 2330, May 1998,

[5] Jack Drobisz, Kenneth J. Christensen, Adaptive Sampling Methods to Determine Network Traffic Statistics including the Hurst Parameter, 23rd. Annual Conference on Local Computer Networks, October 11-14, 1998.

[6] Brewington, B.E. and Cybenko G. (2000), "how Dynamic is the web?" 9th world wide web conference, May 2000.

[7] Li JingChang, "Sampling Investigation and Deduction", the Publisher of China Statistics, 1995. (in Chinese)

[8] Tang XiangNen, Dai JianHua, "Mathematical Statistics", the Publisher of the Mechanism and Industry, 1994. (in Chinese)

[9] Xu baolu, "sampling theory", the Publisher of the BeiJing University, 1982. (in Chinese)

The title: "**Traffic Behavior Analysis with Poisson Sampling on High-speed Network**"

Author's name: Guang Cheng （程光）, Jian Gong （龚俭）

Author's organization: Department of Computer Science and Technology, Southeast University

东南大学计算机系

Correspondence Address: NanJing 210096, P.R.China

中国江苏省南京市，邮政编码：210096

Telephone number: (8610) 025-3794341-210

E-mail address: gcheng@njnet.edu.cn （程光）， jgong@njnet.edu.cn （龚俭）

The paper is intended to submit to the E-09: "Information Network management" of the CONFERENCE E: "Progress, Problems and New Trends in Information Network"