Estimating Properties of Flow Statistics using Bootstrap

Stenio F. L. Fernandes, Tatiene Correia, Carlos A. Kamienski, Djamel F. H. Sadok
{sflf, tatiene, cak, jamel}@cin.ufpe.br
Computer Science Center, Federal University of Pernambuco
CP 7851, Cidade Universitaria, Recife, PE, Brazil, 50732-970

Ahmed Karmouch (karmouch@site.uottawa.ca)
School of Information Technology and Engineering (SITE), University of Ottawa
161 Louis Pasteur, P.O. Box 450, Stn A,
Ottawa, Ontario, Canada, K1N 6N5

Abstract

Traffic measurement has been gaining increasing attention of the network community in the last years, due to its application in a variety of important areas, such as traffic engineering and network planning. Much effort has been devoted to passive flow measurement since collecting packet-level information in high speed links makes this process extremely complex and expensive. There are some techniques for dealing with flow statistics in current commercial routers and associated measurement infrastructure. However, even though flow-level information is more compact than packet-level information, transmitting and storing it would still impose a significant burden on the operation of a typical Internet Service Provider (ISP). In this paper, we advocate that only a small portion of the flow records need to be preserved for further processing. We propose the use of the Bootstrap resampling technique for deriving statistical properties from a previously pre-processed sampled set of flows. Our results show that only 10% or less of the original sampled statistics is necessary in order for Bootstrap to reconstruct the main characteristics of the original raw flow records.

Key words

Network Traffic Measurement; Sampling Technique.

1. Introduction

Monitoring backbone network traffic is a mandatory task to manage today's complex Internet Service Providers (ISP) infrastructure. Particularly, computer networking researchers have made great efforts to make the systemic nature of the Internet more comprehensible, based on passive and active measurements. Hence, network measurements are essential for appraising systems performance, identifying and locating problems in high-speed links [11]. Further, measurement information has been widely used by ISPs for short-term monitoring [9], long-term traffic engineering and provisioning [10], and accounting [1].

In order to obtain such information, today's routers offer tools such as NetFlow [5] that provides flow level information about traffic. The main obstacle with the flow measurement approach is its lack of scalability with link speed [3][7]. Therefore, packet-sampling techniques are progressively being used in routers to export statistics of a fraction of the network traffic [4]. One difficulty that arises is how to recover statistics of the original traffic from such partial sampled data through some reliable procedure. Moreover, due to the huge amount of data produced by flow measurement, it is necessary for routers to control the usage of processing resources, network capacity used to transfer data to collectors, and processing and storage costs at the collectors. Likewise, collecting IP packet headers will give rise to an immense amount of data.

This paper analyses the possibility to derive statistical properties of the original traffic stream from the packet sampled flow statistics, using a resampling technique called Bootstrap [13]. The Bootstrap method follows the plug-in principle, which states that given a parameter of interest θ depending on CDF F, estimate it by replacing F by its empirical counterpart obtained from the observed data.

We found such methodology very appealing and we think it could be applied to a number of circumstances. For instance, Bootstrap estimates could be accurately inferred from light and heavy tailed distribution functions. This opens the doors to the possibility of performing a smooth sampling technique and also achieving a high level of accuracy of the Bootstrap estimates related to the original traffic characteristics. Therefore, considering a variety of network traffic profiles, it seemed promising to look at alternative procedures to reduce the volume of the sampled network traffic data.

Using Bootstrap analysis to characterize the statistical properties of data has lately become a useful and widespread tool in a number of research fields[15][16][17] [18]. For instance, Buvat and Riddel [15] proposed the nonparametric bootstrap method to characterize the statistical properties of computed tomography images. White and Racine [16] investigated the use of bootstrap methods for inference using artificial neural networks applied to predictive accuracy in foreign exchange rates. Recently, Lei and Smith [17] presented some results on an empirical analysis of the reliability of nonparametric bootstrap method in assessing the accuracy of sample statistics in the context of software metrics. Liu et al. [18] proposed the use of Bootstrap method in order to predict fine time-scale behavior of network traffic from coarse time-scale aggregate measurements. Therefore, as far as we

know, research works related to applying bootstrap in the computer network field have not been well explored.

The paper is organized as follows. Section 2 presents related work. Section 3 develops the basic theory of the Bootstrap methodology. The next section (Section 4) shows our validation results based on real network traffic. Finally we draw some conclusions and present suggestions for future work in Section 5.

2. Related Work

There are a number of recent works related to the problem of packet sampling and recovering statistics of the original traffic from sampled data. Some of them focus only on the problem of sampling inside routers whereas others are more interested in resource utilization and analysis in a measurement infrastructure [1][3][7][14].

Estan and Varghese [7] proposed two scalable algorithms for identifying large flows named "sample and hold" and multistage filters. They found out such algorithms to be highly efficient because they would take a constant number of memory references per packet and use a small amount of memory.

In [3] Duffield et al. presented approaches to accurately infer the distributions of flow lengths in the original Internet traffic based on the flow statistics formed from sampled packet streams. Their main contribution is inferring flow numbers and lengths of the original traffic that escaped thinning process completely. They reasoned that as only sampled flow statistics are available some statistical inference is needed to fully determine the flow characteristics of the original unsampled traffic. Also, in [1] Duffield et al. replaced uniform sampling with size dependent sampling. Hence, this approach allows controlling the rate at which samples are produced.

Similarly, in [2] Duffield et al. determined resource usage, for both construction and transmission of flow statistics, and showed how it depends on the flow's characteristics. Afterward they recovered some detailed statistical properties of the original packet stream from the packet sampled flow statistics.

Traffic Analysis Platform (TAP) was proposed in [8] to support detailed information on network resource usage, such as the relative volumes of traffic using different protocols, traffic matrices or the aggregate statistics of packet and byte volumes and durations of user sessions. TAP relies on a distributed infrastructure and on the use of sampling and aggregation at different measurement locations.

3. Bootstrap

Consider a single homogeneous sample of data, denoted by $y_1,...,y_n$. Let the sample data be outcomes of Independent and Identically Distributed (IID) random variables $Y_1,...,Y_n$ whose Probability Density Function (PDF) and Cumulative Distribution Function (CDF) we shall denote by f and F, respectively. The sample is used to make inferences about a population characteristic, generically denoted by θ, using a statistic T whose value is t.

Figure 1 is a schematic diagram of the bootstrap method as it applies to one-sample problems [13]. On the left frame, an unknown distribution F has generated the observed data $y = (y_1, y_2 \ldots, y_n)$. We have calculated a statistic of interest from y, $\theta = s(y)$, and wish to know something about $\hat{\theta}'s$ statistical behavior (e.g., its standard error $se_F(\hat{\theta})$). On the right frame, the empirical distribution gives bootstrap samples $y^* = (y_1^*, y_2^*, \ldots, y_n^*)$ by random resampling, from which we calculate bootstrap replications of the statistic of interest, $\hat{\theta}^* = s(y^*)$.

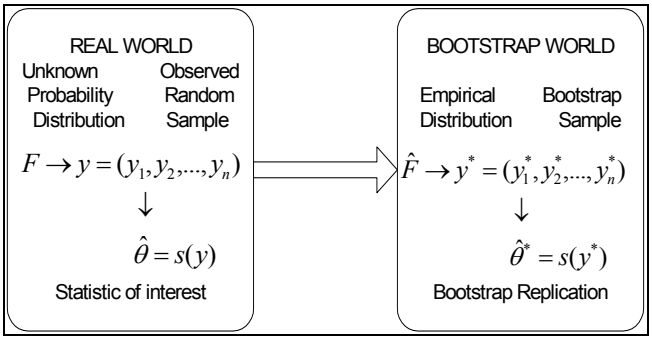

Figure 1 - A schematic diagram of the bootstrap: one-sample problem.

The Bootstrap methodology suggests that if we presume that the sample $y = (y_1, y_2 \ldots, y_n)$ itself represents the underlying distribution we could resample from *y* several times (i.e., the number of bootstrap replicas, *nb*) and compute the statistic $\hat{\theta}^*$ for each of these resamples. Hence, we get a bootstrap distribution of $\hat{\theta}^*$, and for which a confidence interval for $\hat{\theta}$ could be derived. Some authors argue that Bootstrap works fine for the simple reason that, if you have no knowledge about the samples, the data itself will be the best possible approximation of the underlying probability distribution from which they came.

There are two types of Bootstrap techniques: parametric and nonparametric procedures. When we have a particular probability distribution model, with parameters that fully determine f, such a model is termed *parametric* Otherwise, the statistical analysis is nonparametric, and it relies only on the fact that the random variables Y_j are IID [12].

We present now one example to reveal the power of the Bootstrap method. We utilized one dataset with 10,000 samples generated from a Weibull PDF. The 'dataset 1' were deterministically sampled, drawing 1 to N samples (N = 1000, 100, 10). The resulting sampled data will have sample lengths of size n = 10, 100 and 1000. Figure 2 presents the Quantile-Plots from the 'dataset 1' (*nb*= 500). One should notice that even considering the length of the sampled data as low as 1% of the original data size, the

technique could precisely mimic the original quartiles for the Weibull distribution.

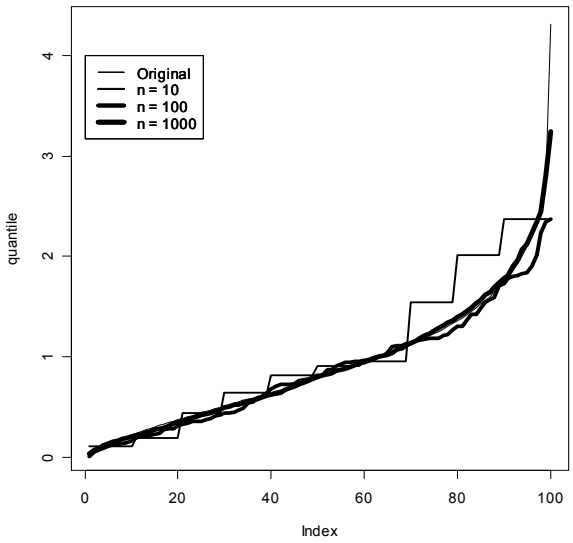

Figure 2 - Quantile-Plot – 'dataset 1'

4. Results

In this section we present a numerical evaluation of the Bootstrap technique tackling some flow level statistics. The passive measurements (raw data) used to illustrate the Bootstrap methodology came from NLANR [6]. We pre-processed such traces to get some metrics, namely the flow lengths (in seconds) and volume (in bytes).

Table 1 and Table 2 present the first and second-order statistics and their respective biases for three sample lengths considered (n = 15, 145, 1447, which correspond to 0.1%, 1% and 10% of the original trace, respectively). These sample lengths are the result of deterministic sampling performed in the pre-processed traces. The original sample dataset comprised approximately 15000 individual (flow-level metrics) records. We also kept the number bootstrap replications (500 and 1000) for both 'duration' and 'volume' datasets.

Table 1 - Descriptive measurements for the 'time' dataset; $n = 15, 145, 1447$ and $nb = 500, 1000$. Original Mean (for Duration=35.06s; for Volume=0.14MBytes)

# of Bootstrap Replicas	*Sample Length*	*Per-flow Duration (s)*		*Per-flow Volume (MBytes)*	
		Mean	*Bias*	*Mean*	*Bias*
500	15	28.74	-6.32	0.01	-0.13
	145	36.11	1.05	0.20	0.06
	1447	35.23	0.16	0.15	0.01
1000	15	29.99	-5.07	0.01	-0.13
	145	36.17	1.11	0.22	0.07
	1447	35.24	0.18	0.14	0.00

Table 2 - Descriptive measurements for 'volume' dataset; $n = 15, 145, 1447$ and nb = 500,1000 . Original Variance (for Duration =1.23e3; for Volume=2.23e12)

# of Bootstrap Replicas	*Sample Length*	*Per-flow Duration*		*Per-flow Volume*	
		Variance (x1e3)	*Bias*	*Variance (x1e10)*	*Bias (x1e10)*
500	15	1.17	-614	1.73	-223
	145	1.23	-4.17	167	-56.3
	1447	1.23	2.35	131	-92.7
1000	15	1.22	-17.1	0.0816	-223
	145	1.22	-6.42	173	-49.9
	1447	1.23	2.01	132	-91.0

In our experiments, we observed similar behavior to the simulation of section 2. Recall that the mean value for per-flow metrics presented in Table 1 refers to the average of flow duration and sizes. We could draw several conclusions from the results. First, if we focus on bias, we could state that increasing the sample length implies in diminishing the value of this metric for both datasets 'duration' and 'volume'. For instance, considering nb = 1000 and n = 145, the bias is 1.11. If we augment the sample length to n = 1447, then the bias decreases to only 0.18. Second, as far as bootstrap replications are concerned, we observed that as nb increases from 500 to 1000, the results get better in most cases. For instance, for nb = 500 and n = 15, the bias for the variance is –614. Increasing twice the number of bootstrap replications, it reaches only 2.35, as shown in Table 2.

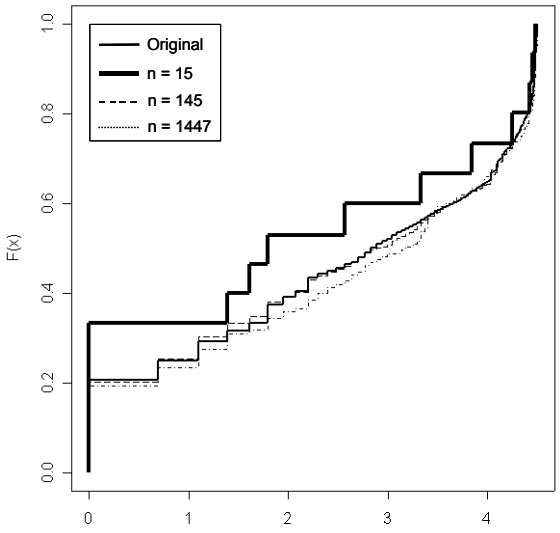

Figure 3 - Empirical Cumulative Distribution Function - 'duration'.

Figure 3 and Figure 4 present the ECDF from the 'duration' (in seconds) and volume datasets, respectively. We set the number of Bootstrap replications (nb) to 500.

One should observe that even with the length of the sampled data as low as 1% of the original data size, the Bootstrap technique could precisely mimic the original raw pre-processed data. Furthermore, if we gather records through 1 in 10 sampling from the original data the resulting ECDF remains indistinguishable.

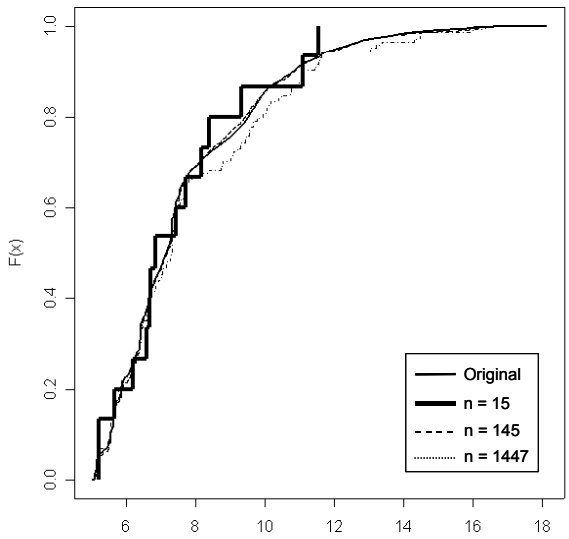

Figure 4 - Empirical Cumulative Distribution Function - Volume.

5. Concluding Remarks

Motivated by the concerns of network operators in large ISP backbones that point out to an ever increasing huge amount of data produced by the passive measurement infrastructure, this paper undertake the problem of reducing such data volume without missing crucial statistical properties. We relied on the Nonparametric Bootstrap technique, which is a resampling procedure. Due to its flexibility, we propose Bootstrap to be used as a technique to reduce data volume either in routers or in a post-processing element.

On applying the methodology in real network traffic measurements, this paper showed that one could use Bootstrap to infer some general characteristics of the network traffic distribution. This paper points out that after executing a short pre-processing in the raw data and extracting some metrics from traces (e.g., flow size and duration), it is necessary to store (in case of Data Warehouse) or transmit (in case of routers) only 10% of the original sampled statistics, in order for Bootstrap to reconstruct its main properties. Our results showed that we could precisely recover the ECDF.

There are several possibilities for future advances. In particular, we would like to analyze the combination of the Bootstrap methodology with the size dependent sampling [3] or inverted sampling [4] techniques. We also wish to verify the computational overhead on deploying Bootstrap in a passive measurement infrastructure.

ACKNOWLEDGMENTS

The authors thank CAPES (BEX0016/04-7) for financial support and Guthemberg Silva for support with the datasets.

References

[1] N.G. Duffield et al., “Charging from Sampled Network Usage”, ACM SIGCOMM IMW 2001, S. Francisco,CA.

[2] N.G. Duffield, et al., “Properties and Prediction of Flow Statistics from Sampled Packet Streams”, ACM SIGCOMM IMW 2002, Marseille, France.

[3] N.G. Duffield, et al., “Estimating Flow Distributions from Sampled Flow Statistics”, ACM SIGCOMM, Karlsruhe, Germany, 2003.

[4] N. Hohn and D. Veitch. “Inverting Sampled Traffic”, ACM Internet Measurement Conf. 2003, Florida, USA.

[5] CISCO NetFlow, http://www.cisco.com – accessed September 16, 2004.

[6] NLANR Measurement and Network Analysis Group, http://moat.nlanr.net/, accessed September 16, 2004.

[7] C. Estan and G. Varghese. “New Directions in Traffic Measurement and Accounting”, ACM SIGCOMM 2002, Pittsburgh, Pennsylvania, USA.

[8] N. Duffield & C. Lund. “Predicting Resource Usage and Estimation Accuracy in an IP Flow Measurement Collection Infrastructure”. ACM Internet Measurement Conference, October 27–29, 2003, Florida, USA.

[9] Ratul Mahajan et al., “Controlling High Bandwidth Aggregates in the Network”, ACM SIGCOMM CCR, Volume 32, Issue 3 (July 2002), Pg 62–73.

[10] K. Papagiannaki, et al., "Long-Term Forecasting of Internet Backbone Traffic: Observations and Initial Models". IEEE INFOCOM. San Francisco. March 2003.

[11] K. Papayanaki, N. Taft, C. Diot. "Impact of Flow Dynamics on Traffic Engineering Design Principles". IEEE INFOCOM 2004. 7-11 March. Hong-Kong.

[12] A. C. Davison & D. V. Hinkley. “Bootstrap Methods and Their Application”. NY: Cambridge U. Press, 1997.

[13] B. Efron and R. J. Tibshirani. “An Introduction to the Bootstrap”. New York: Champman & Hall, 1993.

[14] A. Kumar, J. Xu, L. Li, and J. Wang, “Space-Code Bloom Filter For Efficient Traffic Flow Measurement", ACM SIGCOMM Internet Measurement Conf., 2003.

[15] I. Buvat.; C. Riddell. “A Bootstrap Approach for Analyzing the Statistical Properties of SPECT and PET Images”. IEEE Nuclear Science Symposium Conference Record, Vol. 3, Nov. 2001.

[16] Halbert White and Jeffrey Racine. “Statistical Inference, The Bootstrap, and Neural-Network Modeling with Application to Foreign Exchange Rates”. IEEE Trans. on Neural Networks, V.12, No. 4, July 2001.

[17] Skylar Lei and M.R. Smith. “Evaluation of Several Nonparametric Bootstrap Methods to Estimate Confidence Intervals for Software Metrics”. IEEE Trans. on Software Engineering, Vol. 29, Issue: 11, Nov. 2003.

[18] Chuanhai Liu, S. Vander Wiel and Jiahai Yang. “A Nonstationary Traffic Train Model for Fine Scale Inference from Coarse Scale Counts” IEEE Journal on Selected Areas in Communications, Vol. 21, Aug. 2003.