This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the WCNC 2008 proceedings.

On the Accuracy of Sampling Schemes for Wireless Network Characterization

T. Bheemarjuna Reddy, B. S. Manoj, and Ramesh Rao
Department of Electrical and Computer Engineering, University of California San Diego, CA 92093
btamma@ucsd.edu, bsmanoj@ucsd.edu, and rrao@ucsd.edu

***Abstract*— Wireless network characterization is an important task in next generation wireless networks. In order to achieve efficient wireless network characterization, accurate sampling strategies are required. The relative performance of different sampling strategies for assessing various wireless network traffic metrics is significant due to the complexity and expense involved in the collection, storage, and analysis of all the traffic generated in the wireless medium. Since the spectrum used for most wireless networks, especially those based on IEEE 802.11 standards, is divided into several channels, the existing count-based sampling methods demand continuous capture on each channel for selecting the desired packets of interest. Continuous capturing makes the cost of monitoring infrastructure very expensive and hence count-based sampling methods are not scalable. However, the time-based sampling methods which were considered inaccurate in wired network characterization, appear to offer a cost-effective and scalable solution by reducing the cost of resources necessary to accurately characterize the wireless medium. For example, the use of time-based sampling enable us to make use of a single wireless interface for accurately sampling multiple channels. However, in order to achieve this, we need to identify the right set of parameters for time-based sampling. This paper presents a study of the performance of various time-based sampling methods in answering questions related to their use in wireless network traffic characterization. We simulate time-based sampling traces at a variety of granularities using a complete packet trace (*i.e.,* parent population) captured in a campus wireless network environment that aggregates traffic from a large number of nodes. From our analysis using Chi-Square test, we found that the Timer-driven Time-based sampling is more accurate than Count-driven Time-based sampling for both systematic and stratified sampling schemes.**

I. Introduction

Wireless network characterization, especially in an IEEE 802.11 network, is useful for many applications such as network traffic characterization, capacity planning, network management, optimizing deployment of Access Points (APs), detecting network anomalies, and cognitive networking. Cognitive networks [1], [2] gather, compact, analyze, and repositorize large amounts of spatio-temporally tagged wireless network data as well as users network experience information in order to better optimize the network resource management. With the advent of high-speed 802.11 a/g/n technologies, the cost of collection, storage, and analysis of all the traffic generated in the air across various channels becomes too expensive. As a scalable means to monitor wireless network traffic, packet sampling has attracted much attention from both industrial and research communities. Sampling is a form of passive traffic measurement. In sampling, not all packets are measured, but only a selected fraction based on the sampling method employed and the sampling parameters chosen. Hence sampling methods reduce the measurement data. The data reduction not only reduces the bandwidth consumed in transmitting the measurement data to the collection point, but also decreases the cost incurred for analysis and storage of the data. The deployment of sampling methods aims at estimating some specific characteristics of the parent population (*i.e.,* the complete network traffic) at a lower cost than a complete census would demand. Sampling trades off the opposing goals of controlling estimation accuracy and measurement costs. Both the IETF (Internet Engineering Task Force) working groups, IPFIX (IP Flow Information Export) and PSAMP (Packet Sampling), have recommended the use of packet sampling. Count-based systematic sampling method such as "1 out of N packets" is a popular sampling design employed in Cisco and Juniper routers.

Sampling methods can be characterized by the sampling algorithm (which describes the basic process for selection of packets from each sampling interval) and the trigger type used for starting the packet capture. Based on the sampling algorithm, there are three main classes of sampling methods: systematic sampling, stratified random sampling, and simple random sampling [3]. For each class, one can use either packet counts or timers to trigger the selection of packets for inclusion in a sample. In systematic sampling packets for inclusion in a sample are selected deterministically from each sampling interval. Stratified random sampling involves selecting packets randomly from each sampling interval, whereas in the case of simple random sampling packets are selected randomly from the parent population. Based on trigger type used for starting the packet capture, sampling methods can be broadly classified into count-based and time-based sampling.

In count-based sampling, packet count triggers the start of a sampling interval. Length of a sampling interval is called sampling period or cycle. Here sampling period is defined by the number of packets. Sampling duration or length is defined as the number of packets selected for inclusion in the sample from each sampling interval. An example of systematic count-based sampling is to select every n^{th} packet in the packet stream. An example of stratified count-based sampling is to randomly select a packet in every n packets. For both examples sampling period or cycle is n packets and sampling duration or length is one packet.

In time-based sampling, timer triggers the start of a sam-

TABLE I
TIME-BASED SAMPLING SCHEMES

Sampling Method	Sampling Algorithm	Trigger Type
SCT	Systematic	Count-driven Time-based
STT	Systematic	Timer-driven Time-based
SRCT	Stratified Random	Count-driven Time-based
SRTT	Stratified Random	Timer-driven Time-based

pling interval or sampling period. Hence sampling period is defined by a timer. But sampling duration can either be timer-driven or count-driven. Hence time-based sampling can be further classified into Timer-driven time-based sampling and Count-driven time-based sampling. An example of systematic timer-driven time-based sampling is to capture all packets arriving in first 1 sec of every 11 sec. An example of systematic count-driven time-based sampling is to sample a packet every 11 sec. For both examples sampling period or cycle is 11 sec, but sampling durations or lengths are 1 sec and one packet, respectively. Various time-based sampling schemes are given in Table I.

A. Challenges in Wireless Network Characterization

There exist several challenges offered by the wireless network environment. Some of which are discussed here.

Multiple Channels: The traffic in wireless environment, especially in the ISM band, is spread across a number of channels where a single monitoring network interface can typically access only one channel at a time. In order to characterize the traffic across all the channels, the monitor node either has to have one interface tuned to operate in each channel or employ a multi-channel sampling strategy in which a single monitoring interface need to be switched across the channels. In the first case, the monitor node is likely to be very complex or infeasible due to the presence of a large number of channels in the network. For example, the 802.11 b/g and 802.11 a-based wireless networks have about 11 and 13 channels, respectively. Hence a multi-interface monitor node needs to use 24 interfaces to collect the complete traffic from all channels. Moreover, transmission, storage, and analysis of captured packets from a large number of simultaneous channels may be problematic, with each capture node capable of capturing almost 600 Mbps of traffic (assuming 11 channels at maximum 802.11 link-layer rates: 54 Mbps). Therefore, such a multi-interface complete capture can be very expensive and not scalable for characterizing large scale wireless networks; therefore, scalable multi-channel traffic sampling strategies assume an important role in wireless networks.

Traffic characteristics of wireless environment: Unlike wired networks, the wireless environment has a number of unique characteristics that would affect the accuracy of the traffic sampling strategy. Some of these characteristics include: (i) presence of aggregate traffic, (ii) location dependent contention, (iii) presence of broken or corrupt packets, and (iv) multi-rate data transmissions. The presence of aggregate traffic in the local wireless environment makes the traffic modeling and prediction a very challenging task. This is because the traffic characteristics do not follow well known statistical distributions such as poison or exponential. The wireless traffic behavior is highly location dependent and therefore, the traffic characterized for a given spatio-temporal coordinates may not be suitable even for a slight change in the coordinates. The presence of broken or corrupt packets can contribute to the consumption of channel resources exactly similar to a successful packet, however, such packets cannot be easily received by commercial wireless network interface cards and hence they may not be included in the traffic trace. Thus the inability of the monitor node to take into account such packets can lead to erroneous characterization. Above all, unlike in wired networks, the wireless environment offers multiple transmission rates. For example, the IEEE 802.11b can operate in several transmission rates such as 1 Mbps, 2 Mbps, 5.5 Mbps, and 11 Mbps. Therefore the channel resources consumed by a 1Mbps packet is several times more than the channel resources consumed by the same length packet transmitted at 11 Mbps. Unlike the wired network traffic characterization, the wireless traffic characterization can be affected by unknown interference sources as well. For example, the presence of Bluetooth traffic or microwave traffic can potentially affect the 802.11 b/g network traffic thereby making an erroneous traffic characterization.

Not all the above mentioned problems can be solved by traffic characterization using general purpose wireless network interfaces. In this paper we study the accuracy of multi-channel wireless sampling schemes, which come under the Time-based sampling. Time-based sampling methods are not studied thoroughly in the literature. Most of the conclusions derived on accuracy of these methods are based on count-driven time-based sampling with some fixed sampling period and sampling duration of one packet. In this paper we would like to study the accuracy of various Time-based sampling schemes by varying sampling periods and sampling durations for traffic characterization in wireless networks.

On the data network traffic characterization, Claffy et al. in [5] presented a detailed study of the performance of various data traffic sampling methods for wide area network traffic. They answered a number of questions on wide area network traffic characterization including the sampling accuracy of time and count-based methods for both random and systematic periods of sampling. However they only studied count-driven time-based sampling. Their result mainly pointed to the inappropriateness of using the count-driven time-based techniques because they do not perform as well as the count-based ones. This is due to traffic patterns, *i.e.,* it is well established that Internet traffic is bursty over a range of time scales. Consequently there can be inhomogeneous bursts of many packets with small inter-arrival times. Time-based sampling methods more easily miss these than a count-based method, and estimators built on them have higher variance. However, as mentioned above, implementing a count-based sampling method is very expensive in a multi-channel wireless network environment. Since timer-driven time-based sampling is not studied in the literature, in this paper we would like to study

the accuracy of time-based sampling schemes for wireless network traffic characterization.

Desphande et al. [6] proposed two methods for channel-based sampling in IEEE 802.11 b/g networks. The first method is a timer-driven time-based sampling with a fixed sampling duration (1 sec) and fixed sampling period (11 secs). Their second method adaptively varied the sampling duration as a function of the packet arrivals seen on each channel. However, the important information required, the relation between the sampling duration and the sampling period, is not studied in their paper. From our experiments, we found that not all combination of sampling periods and sampling durations can be used for accurate timer-driven time-based multi-channel sampling.

II. Sampling Methods and Traffic Metrics

We can characterize a sampling method with the following four parameters: sampling algorithm, trigger type, sampling period, and sampling duration. The values of these parameters should be chosen based on the accuracy requirements and the sampling overhead, as well as the characteristics of the parent population and the traffic metrics being measured. In this work, we take three traffic metrics and compare the sampling methods based on how correctly the information can be extracted from the sampled traces. Metrics of interest are: packet size distribution, packet inter-arrival time distribution, and packet data rate distribution. We implemented four time-based sampling methods (refer Table I) at different granularities by varying sampling duration and sampling period. Our goal is to study the effect of certain sampling parameters on the integrity of the resulting samples.

A. Evaluation of Discrepancy

To determine and quantify the performances of various sampling methods, we need discrepancy measurements to gauge how close the distributions given by the sampling methods are, compared to the actual distributions of the parent populations. Pearson's chi-square test statistic is a measure of the discrepancy between the observed and expected counts within a set of bins which span the range of data. It is defined as:

$$X^2 = \sum_{i=1}^{N} \frac{(O_i - E_i)^2}{E_i} \quad (1)$$

where N is the number of bins, O_i is the number of observations found in the i^{th} bin of the sampled data, and E_i is the number of observations expected in the i^{th} bin based on the parent population model. The sampling distribution of X^2 is approximately the chi-square distribution where the number of degrees of freedom equals the number of bins minus one. This approximately improves as the number of counts in each bin increases, and is generally adequate if each bin has at least five expected counts. This statistic is the basis for the chi-square test, which uses the chi-square distribution to test hypotheses at specified significance level about the goodness of fit between the parent population model and the sampled data. Given a sampled data generated from one of the sampling methods, we test the null hypothesis that the distribution of sampled data agrees with or "fits" the distribution of parent population. For this we need to find the values of test statistic, X^2, and the critical value, CV. The CV separates the critical region (the set of values of the test statistic that cause us to reject the null hypothesis) from the values of the test statistic that do not lead to rejection of the null hypothesis. It is defined as $CV = Chi2Inv(1-\alpha, df)$, where $Chi2Inv$ is the inverse chi-square cumulative distribution function, α is the significance level (the probability that the test statistic will fall in the critical region when the null hypothesis is actually true), and $df = N - 1$ is the degrees of freedom. To test the null hypothesis the upper tail of chi-square distribution is used as the critical region. If $X^2 \geq CV$, we reject the null hypothesis (i.e., the distribution of sampled data does not fit the distribution of parent population). Otherwise we fail to reject the null hypothesis.

III. Experiments and Results

A. Monitoring Infrastructure

All of our wireless traffic monitoring activity take place within UCSD division of CALIT2, a large six-story building. Avaya APs provide production wireless service, configured for 802.11 b/g service. Further there exists some experimental mesh networks on the sixth floor. Between and among production APs located on 4th and 6th floors, we have deployed 12 CalNodes (6 on each floor). Each CalNode consists of a Soekris Engineering net4521 system board with two 100 Mbps Ethernet interfaces and one Ubiquity 802.11 a/b/g cardbus wireless interface based on Atheros AR5213 chipset with external antenna connectors. Each CalNode runs a version of Voyage Linux with kernel 2.6.x and uses the open source MadWiFi driver for driving the Atheros-based wireless interfaces. In order to report additional information about the packet currently being captured, the MadWiFi driver generates the prism monitoring header of size 144 bytes and adds it to the packet. The prism monitoring header contains received signal strength indicator (RSSI), capture device, channel, data rate, and other signal/noise quality information. Each CalNode is connected to the campus intranet via one of the Ethernet interfaces.

Using the capture-to-file functionality of the open source tcpdump packet sniffer, the CalNodes create capture files and remit these to a central repository Dell PowerEdge 1900 server (two Dual Core Intel Xeon processors operating at 2 GHz with 4 GB RAM and 4.2 TB of storage) via FTP. To further reduce the storage cost, we configured tcpdump to capture only the first 250 bytes of each sampled packet. This is a reasonable solution, since TCP header and other protocol headers are located at or near the start of the packet. At the repository, a modified version of tcpdump is employed to read the capture file to extract prism monitoring header fields and header field values from the MAC through transport layers of the TCP/IP protocol stack. These values are stored in a MySQL database

TABLE II

SUMMARY STATISTICS FOR DISTRIBUTION OF PACKET SIZES, DATA RATES, AND INTER-ARRIVAL TIMES FOR THE COMPLETE PACKET TRACE.

Distribution	Min	25%	Median	75%	Mean	Max	StdDev	Skew	Kurtosis
Packet sizes (Bytes)	154	214	222	222	220.62	1676	71.66	12.36	205.58
Data Rates (Mbps)	1	2	2	2	2.19	54	2.29	14.38	255.58
Inter-arrival Times (Microsec)	57	1000	1000	16000	13089	626000	22919	2.90	31.62

from which they can be queried using MATLAB for analysis purposes.

To simply our sample collection process, we configured a CalNode to do continuous packet capture on one particular channel for 24 hours. This trace is collected on the 27 September 2007 on channel 11 alone. It has 6,544,722 packets in it and the corresponding MYSQL table is of size 1.28 GiB. We treat this trace as our parent population data set and generate different sample traces by varying sampling duration and sampling period for various Time-based sampling methods. Table II quantifies the parameters of packet arrival rate, mean packet data rate, and mean packet size distributions for the population packet trace.

B. Results

We conduct chi-square goodness of fit test for the various samples generated by SCT, SRCT, STT, and SRTT sampling schemes by varying sampling parameters like Sampling Duration (SD) and Sampling Period (SP). In this study our target distributions are inter-arrival times, packet sizes, and data rates. For each of the sampling schemes, we generate samples at different level of granularity. For each of these sampled data sets, we test the null hypothesis that the distribution of sampled data agrees with or "fits" the distribution of parent population.

1) Bin selection and Significance level: Calculation of chi-square test statistic and the corresponding critical value requires the selection of bins, or ranges, in which to group the data sets and the significance level, α, for rejecting the null hypothesis. But chi-square test is very sensitive to N and α. According to [8], for large sampled data sets, the number of bins (N) should be in the range: $1.88M^{2/5} < N < 3.76M^{2/5}$, where M is size of the sampled data set. So in our study, we set $N = 2M^{2/5}$ and then divided the range of data values into equal size bins. The significance level of 0.05 is used for rejecting the null hypothesis in our goodness of fit tests.

2) Distribution of Inter-arrival Times: Figures 1 and 2 show chi-square test scores for various samples obtained from SCT, SRCT, STT, and SRTT sampling schemes. In this experiment we kept SP constant at 11 seconds and varied SD to obtain various samples. For SCT and SRCT sampling schemes SD is defined as the number of packets to capture for inclusion in the sample during SP. But for STT and SRTT schemes SD is a timer; all packets that arrive before expiry of the timer are included in the sample. Figures also show critical values used for rejecting the null hypothesis. Figure 1 shows that count-driven time-based sampling methods (SCT and SRCT) fail the test for all different values of sampling duration (i.e., null hypothesis of distribution of sampled inter-arrival times fits the distribution of population of inter-arrival times is rejected).

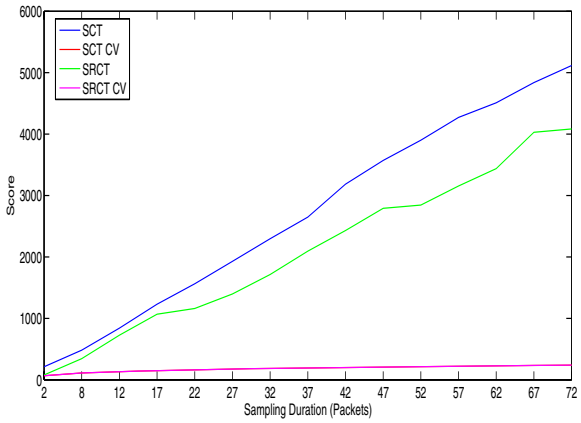

Fig. 1. Goodness of fit test scores of SCT and SRCT schemes as a function of sampling duration for packet inter-arrival time distribution (sampling period=11s).

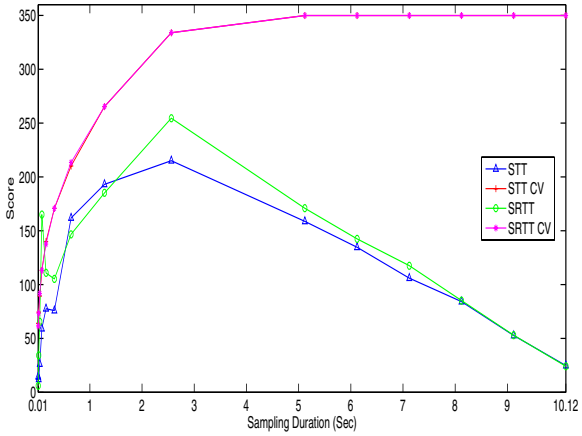

Fig. 2. Goodness of fit test scores of STT and SRTT schemes as a function of sampling duration for packet inter-arrival time distribution (sampling period=11s).

Figure 2 shows the scores of samples obtained from timer-driven time-based schemes, STT and SRTT. The samples are collected at exponentially increasing sampling durations. Unlike count-driven time-based schemes seen above, STT and SRTT schemes pass the test for almost all sampling durations ranging from 10 ms to 10240 ms. As shown in the figure, two samples of SRTT scheme fail the test for smaller sampling durations. We also plotted the empirical Cumulative Distribution Function (CDF) of inter-arrival times

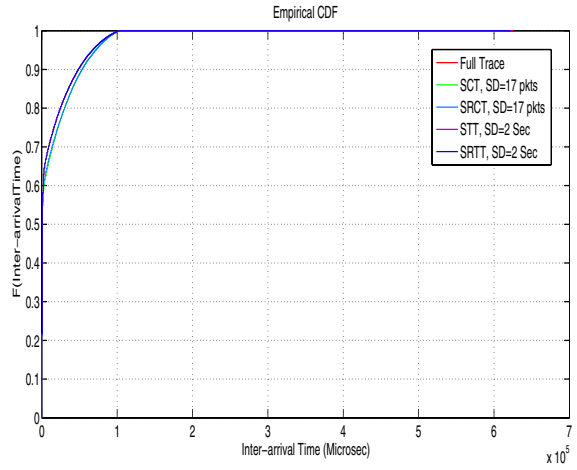

Fig. 3. Cumulative distribution of inter-arrival times for Full packet trace and various sampling schemes (sampling period=11s).

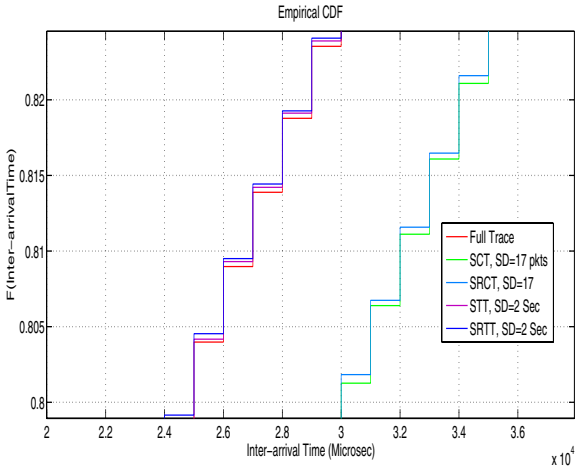

Fig. 4. Zoom-in of Figure 3.

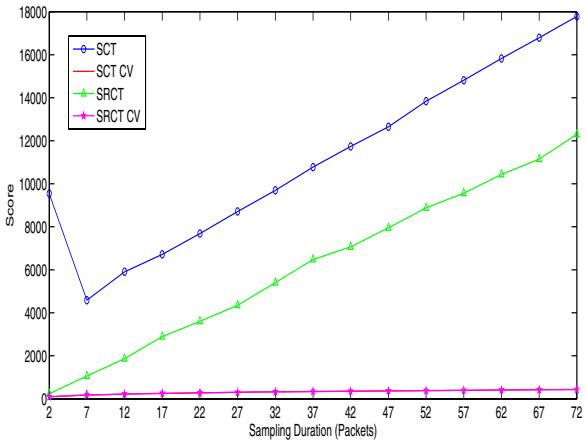

Fig. 5. Goodness of fit test scores of SCT and SRCT schemes as a function of sampling duration for packet size distribution (sampling period=11s).

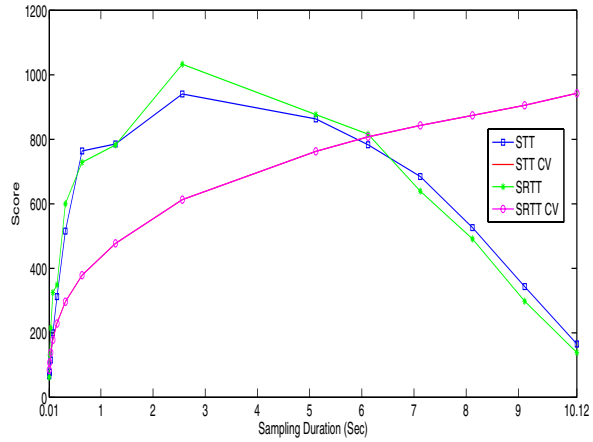

Fig. 6. Goodness of fit test scores of STT and SRTT schemes as a function of sampling duration for packet size distribution (sampling period=11s).

for the full packet trace and various samples obtained from count-driven time-based and timer-driven time-based schemes (refer Figures 3 and 4). Since SRCT and SCT samples fail the test (refer Figure 1), CDF of these samples are not close to the CDF of full packet trace (refer enlarged Figure 4). Hence timer-driven time-based sampling schemes are more accurate than count-driven time-based sampling schemes for varying sampling granularities for the distribution of packet inter-arrival times. Between systematic and stratified random, systematic sampling schemes slightly perform better.

3) Distribution of Packet Sizes: Figures 5 and 6 show chi-square test scores for various samples obtained from SCT, SRCT, STT, and SRTT sampling schemes. In this experiment also we kept SP constant at 11 seconds and varied SD to obtain various samples. Figure 5 shows that count-driven time-based sampling methods (SCT and SRCT) fail the test for all different values of sampling duration. Like for the distribution of inter-arrival times, all samples generated using SCT and SRCT sampling schemes completely fail in representing the characteristics of packet size distribution. Hence as reported in [5] count-driven time-based sampling schemes are uniformly less accurate for characterizing network traffic. No given sampling parameters fit for all distributions, the sampling duration and sampling period should be chosen cautiously based on the distribution of interest. Figure 6 shows that timer-driven time-based sampling methods (STT and SRTT) pass the test for some sampling durations and hence timer-driven time-based sampling is good for representing the characteristics of packet size distribution. Interestingly samples of timer-driven time-based sampling scheme pass the test for both extremely small values of sampling durations (10ms and 20ms, refer Figure 7) and extremely large sample sizes (sampling durations greater than seven seconds). Between systematic and stratified random, systematic schemes slightly perform better. Figures 8 and 9 show the CDF of packet sizes for the full packet trace and various samples obtained from count-driven time-based and timer-driven time-based schemes. We also studied the effect of sampling schemes on the distribution of packet data rates. There also count-driven time-based sampling

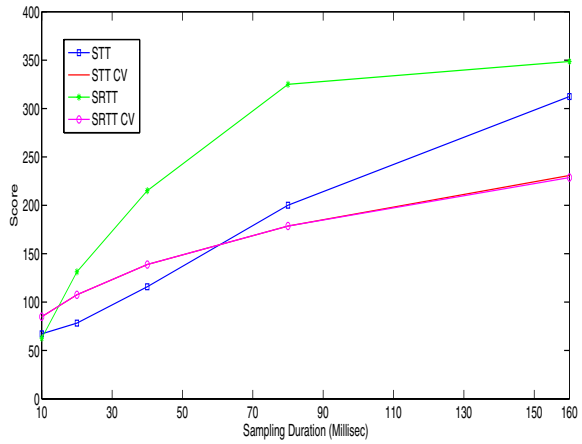

Fig. 7. Zoom-in of Figure 6.

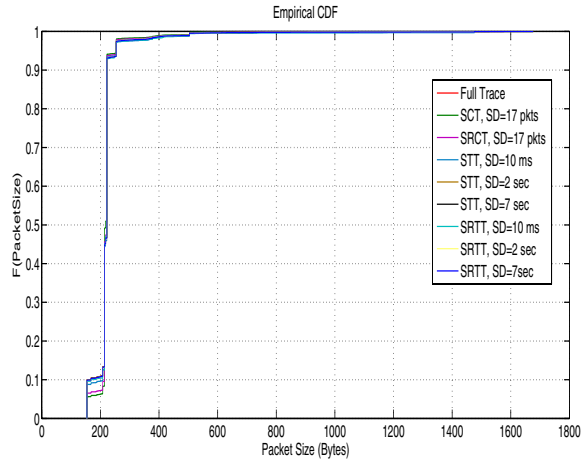

Fig. 8. Cumulative distribution of packet sizes for Full packet trace and various sampling schemes (sampling period=11s).

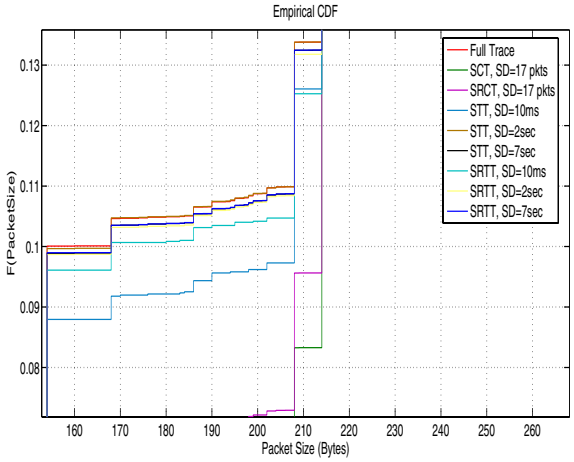

Fig. 9. Zoom-in of Figure 8.

schemes completely fail in representing the characteristics of data rate distribution. We also conducted similar experiments by varying channel number for several working days in our UCSD campus. Results are very similar to what we presented in this paper. Due to lack of space, we are not including those results in this paper.

IV. Conclusions

Wireless network traffic characterization is the main approach using which we can estimate the network performance experience. The presence of multiple channels, multiple data rates, and location dependent contention are some of the issues that affect the feasibility of wireless network traffic characterization strategies. In this paper we studied the performance of various time-based sampling methods in answering questions related to their use in wireless network traffic characterization. From our analysis using Chi-Square goodness of fit test, we found that the Timer-driven Time-based sampling is more accurate than Count-driven Time-based sampling for both systematic and stratified sampling schemes. Between systematic and stratified random, systematic sampling schemes slightly perform better. Like for the distribution of inter-arrival times, count-driven time-based sampling schemes completely fail in representing the characteristics of packet size and data rate distributions. No given sampling parameters fit for all distributions, the sampling duration and sampling period should be chosen cautiously based on the distribution of interest. Finding the characteristics relation between the CDF of the data and its relation to the Chi-square test score is an interesting future work that can throw light on the characteristic features of data that can be used for sampling decisions.

ACKNOWLEDGMENT

This work was supported by NSF sponsored projects, at University of California San Diego, CogNet and Rescue (award numbers: 0650048 and 0331690).

References

[1] R. W. Thomas, D. H. Friend, L. A DaSilva, and A. B. MacKenzie, "Cognitive Networks: Adaptation and Learning to Achieve End-to-end Performance Objectives," *IEEE Communications Magazine*, vol. 44, no. 12, pp. 51-57, December 2006.

[2] B. S. Manoj, M. Zorzi, and R. R. Rao, "Architectures, Protocols, and Analytical Approaches for Next Generation Cognitive Networking," *Book Chapter in Cognitive Wireless Networks: Concepts, Methodologies and Visions*, (Edited by Katz Marcos et al.) Springer, 2007.

[3] Nick Duffield, "Sampling for Passive Internet Measurement: A Review," *Statistical Science*, vol. 19, no. 3, pp. 472-498, 2004.

[4] M. Goyal, R. Guerin, and R. Rajan, "Predicting TCP Throughput from Non-invasive Network Sampling," *in Proc. of IEEE Infocom*, vol. 1, pp. 180-189, June 2002.

[5] K. C. Claffy, G. C. Polyzos, and H.-W. Braun, "Application of Sampling Methodologies to Network Traffic Characterization," *Computer Communication Review*, vol. 23, no. 4, pp. 194-203, October 1993.

[6] U. Deshpande, T. Henderson, and D. Kotz, "Channel Sampling Strategies for Monitoring Wireless Networks," *in Proc. of 4th International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks*, pp. 1-7, April 2006.

[7] A. Balachandran, G. M. Voelker, P. Bahl, and P. V. Rangan, "Characterizing User Behavior and Network Performance in a Public Wireless LAN," *in Proc. of ACM SIGMETRICS*, pp. 195-205, 2002.

[8] B. Schorr, "On the Choice of the Class Intervals in the Application of the Chi-Square Test," *Math. Operations Forsch. u. Statist*, vol. 5, pp. 357-377, 1974.