TIME-BASED SAMPLING STRATEGIES FOR MULTI-CHANNEL WIRELESS TRAFFIC CHARACTERIZATION IN TACTICAL COGNITIVE NETWORKS

Bheemarjuna Reddy Tamma, Manoj B.S., and Ramesh R Rao
California Institute for Telecommunications and Information Technology – UC San Diego, USA
{btamma,bsmanoj,rrao}@ucsd.edu

ABSTRACT

Accurate characterization of wireless network traffic has many applications in military communication networks. However, the network traffic characterization is a challenging task in multi-channel wireless networks. Due to the presence of multiple channels, the existing count-based sampling methods demand continuous capture on each channel for selecting the desired packets of interest. Continuous traffic capture 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, seem to offer a cost-effective and scalable solution.

The contributions of this paper include the following: (i) proposal of a new metric, Relative Proportional Inconsistency (RPI) for measuring the accuracy of sampling schemes, (ii) comparison of various time-based sampling strategies using RPI metric, (iii) studying the effect of sampling parameters on the accuracy of sampling, and (iv) preliminary results on characterization of wireless network data traffic from residential and campus environments.

I. INTRODUCTION

Military operations in recent urban warfare have shown the vulnerability of current military communication networks. Tactical systems operating in such non-traditional battle fronts are facing increased interference, scarcity of bandwidth, attacks from jamming sources, lack of historical traffic information, and lack of proper information collection and sharing mechanisms, there by significantly increasing the death toll of soldiers in such hostile environments. Today, wireless spectrum information is collected using instantaneous measurements onsite at the battlefield. However, in critical situations, the time required for each radio to carry out its individual spectrum sensing activity can lead to unpardonable delays. In addition, the network throughput that the applications receive may not necessarily be proportional to the channel qualities because the end-to-end performance of wireless devices depends on a variety of other factors such as protocol parameters on all the seven layers in the network protocol stack, mobility, and congestion in the network. These scenarios demand sophisticated schemes which characterize wireless network traffic information in a spatio-temporal fashion and play a big role in the network optimization and resource management. The recently emerged reseach area, Cognitive networking [1] which gather, compact, analyze, and reposit 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, also demands sophisticated schemes for wireless network traffic characterization.

There exist several challenges offered by the wireless network environment as discussed in [2]. The most important two issues are arising from the presence of multiple channels and the characteristics of wireless traffic environment. In order to characterize the traffic across all the channels in a multi-channel system, the monitoring device either has to have one interface tuned to operate in each channel or the interface need to be switched across all 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. Moreover, transmission, storage, and analysis of captured packets from a large number of simultaneous channels may be problematic. Therefore, such a multi-interface complete capture solution can be very expensive and not scalable for characterizing large scale wireless networks. The multi-channel sampling scheme is therefore essential.

As a scalable means to monitor wireless network traffic, packet sampling has attracted much attention from industrial and research communities. Sampling is a form of passive traffic measurement, in which not all packets are measured, but only a selected fraction based on the sampling method and parameters associated with the sampling process. Count-based systematic sampling method such as "1 out of N packets" is a popular

978-1-4244-2677-5/08/$25.00

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. 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. We note that implementing a count-based traffic sampling method is very expensive in a multi-channel wireless network environment, as we do not know in advance at what times packets will appear on the channel and packet arrival rates vary across the channels, and therefore it requires one dedicated wireless NIC to sample each channel.

In time-based sampling, timer triggers the start of a sampling 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. The time-driven time-based sampling methods seem to offer a cost-effective and scalable solution by reducing the cost of resources necessary to accurately characterize the wireless traffic. For example, the use of timer-driven time-based sampling enable us to make use of a single wireless interface to sample multiple channels. However, in order to achieve a high sampling accuracy, we need to identify the right set of parameters to be used for time-based sampling.

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

On the data network traffic characterization, Claffy et al. in [4] presented a detailed study of the performance of various sampling methods for wide area network traffic. Their work focused on 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* sampling techniques. Desphande et al. [5] 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 fixed sampling parameters. 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 effect of sampling parameters on the accuracy of sampling, is not studied in their work. Hence, in this paper we would like to study the effect of sampling parameters on the sampling accuracy of various time-based sampling schemes and then present preliminary results on characterization of IEEE 802.11 traffic from residential and campus environments.

II. SAMPLING METHODS AND ERROR 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 four 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, packet data rate distribution, and packet received signal

strength indicator (RSSI) 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. *Error Metrics*

To determine and quantify the performances of various sampling methods, we need an error metric to gauge how close the distributions given by the sampling methods are, compared to the actual distributions of the parent populations. One of the best known metrics is Pearson's Chi-square goodness of fit test, which measures the discrepancy between the observed and expected counts within a set of bins which span the range of traffic metric under investigation. It is defined as [6]:

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

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. This χ^2 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. Unfortunately, the χ^2 statistic is sensitive to the size of sampled data set as well as the number of bins, making it difficult to compare samples of varying sizes [4].

In [7] an alternative metric is presented: the ϕ (phi) coefficient. This metric is free of the influence of the sample size and it is derived from the χ^2 metric as follows: $\phi = \sqrt{\frac{\chi^2}{n}}$, where $n = \sum_{i=1}^{N}(O_i + E_i)$. The value of ϕ coefficient will be zero if the distribution of sampled data coincides exactly with that of parent population. Large values of ϕ will indicate greater discrepancy between distributions of samples and the parent population. However, one important disadvantage of ϕ coefficient is that it cannot quantify the error in the sampling.

Hence, we propose a new metric, Relative Proportional Inconsistency (RPI), for measuring the accuracy of sampling schemes. The RPI metric is defined as follows:

$$RPI = \frac{\sum_{i=1}^{N} \frac{|P_i - S_i|}{P_i}}{N} \quad (2)$$

where P_i (S_i) denotes the proportion of population data (sampled data) found in the i^{th} bin. N denotes the number of bins which span the range of traffic metric under investigation. The RPI metric is derived from the Proportional Consistency (PC) metric [8] where the sampling quality is measured as the level of the proportion consistency, *i.e.,* the consistency between the sample proportion and the population proportion for a set of bins. However, the RPI measures the relative error or proportional inconsistency. The RPI therefore, can provide the quantification of sampling error even when the error is small and cannot be measured easily by the PC metric. Similar to ϕ coefficient, the value of RPI approaches to zero when the distribution of sampled traffic matches that of parent population. Further, RPI is not sensitive to the size of sampled data in comparison to the Chi-square metric. Therefore, it is more preferred metric than both PC metric and Chi-square goodness of fit test.

Bin selection: Calculation of error metrics requires the selection of bins, or ranges, in which to group the population data set for the traffic metric of interest. In our experiments we choose bin edges in such a way that each bin contains at least 1% of data from the population data set.

III. EXPERIMENTAL SETUP

All of our wireless traffic monitoring activity took 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 11 CalNodes (wireless traffic samplers) [9]. Each CalNode consists of a Soekris net4521 system board with one Ubiquity 802.11 a/b/g cardbus wireless interface based on Atheros AR5213 chipset. The wireless interface is configured in *monitor* mode to capture all 802.11 packets in the air. Each CalNode runs on Voyage Linux with kernel 2.6.x and uses the open source MadWiFi driver. Each CalNode is connected to the campus intranet via one of the Ethernet interfaces. CalNodes use the *tcpdump* packet sniffer to create capture files and remit them to the CogNet repository via FTP. To reduce the storage cost, we configured *tcpdump* to capture only the first 250 bytes (144 bytes for Prism monitoring header) of each sampled packet. The CogNet repository is a Dell PowerEdge 1900 server with 7.2 TB of storage. Prism monitoring header fields and other protocol header fields that are extracted from the sampled packets are stored in the CogNet database, implemented using MySQL, from which they can be queried for analysis purposes.

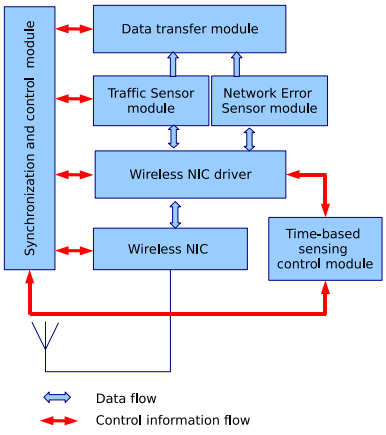

Fig. 1. A schematic diagram of the CalNode.

Figure 1 shows a schematic representation of CalNode. The main components are the time-based sampling control module and the synchronization and control module. The time-based sampling control module controls the time for which the traffic samples from each of the channels are to be collected. This module enables the wireless NIC to scan all the channels in a predefined static or dynamic fashion. On the otherhand, the synchronization and control module performs synchronization of CalNode with CogNet repository using NTP over the wired network; this achieves relative synchronization between CalNodes. Absolute synchronization is obtained by having the Cognet repository synchronized with a public NTP server. Traffic sensor module implements one of Time-based sampling schemes, where as network error sensor module gathers additional statistics like CRC errors and PHY errors.

To get samples at different granularities for various Time-based sampling schemes and measure their sampling accuracy using RPI metric, we need population (*i.e.,* complete) packet traces. Therefore we configured a couple of CalNodes to do continuous packet capture on one particular orthogonal channel for four weeks. We treat these traces as our parent population data sets and generate different sample traces by varying sampling duration and sampling period for various Time-based sampling methods.

IV. Performance Results

We compute RPI and ϕ coefficients of 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, RSSIs, and data rates. In the graphs all performance results are shown with 95% confidence intervals.

A. *Comparison of Sampling Schemes*

Figures 2 and 3 show ϕ coefficients and RPI scores for various samples obtained from STT and SRTT sampling schemes. In this experiment we kept SP constant at 11 seconds and varied SD from 100 ms to 11 secs to obtain various samples from the parent packet size trace. As SD increases (i.e., sample size increases) values of both error metrics gradually decreased. In addition, when the SD goes beyond 1 sec, the variation in the error decreases. The trends for both ϕ coefficient and RPI score are the same. In addition, the RPI can quantify the relative error whereas the ϕ coefficient cannot be easily translated to equivalent error quantity. For the RPI, the mean sampling error is less than 1.5% for the sample generated with SD of 1 sec. In other words, even samples generated with SD of 1 second are closely matching the distribution of parent population. In order to reduce the cost of wireless traffic sensing, we would like to reduce the value of SD to a maximum extent without sacrificing sampling quality. Even though there are 11 channels to be monitored for traffic in IEEE 802.11 b/g spectrum, when we employ timer-driven time-based sampling with sampling parameters SD=1 sec and SP=11 sec, one wireless interface is sufficient to monitor the wireless medium. From figures we can also observe that the performance of STT and SRTT sampling schemes are identical. However, as SRTT requires generation of pseudo-random numbers it is more expensive to implement in real sampling systems. In the following we only plot RPI scores as both RPI and ϕ metrics are exhibiting similar trends.

In Figure 4 we show RPI scores of count-driven time-based schemes, SCT and SRCT. For these 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. As shown in figure the performance of SCT and SRCT schemes are identical. But when we compare these schemes against timer-driven time-based schemes (STT and SRTT), RPI scores are very low for timer-driven time-based schemes than count-driven time-based sampling schemes (SCT and SRCT). This is because in count-driven time-based schemes, sample size is fixed (as SD is defined in terms of packet count) and does not grow linearly with traffic load in the network. But in the case of STT and SRTT schemes, sample size grows

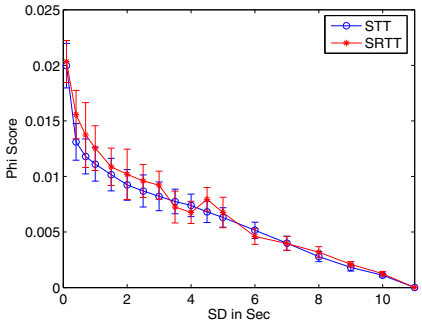

Fig. 2. Phi coefficients of STT and SRTT schemes vs SD for packet size distribution.

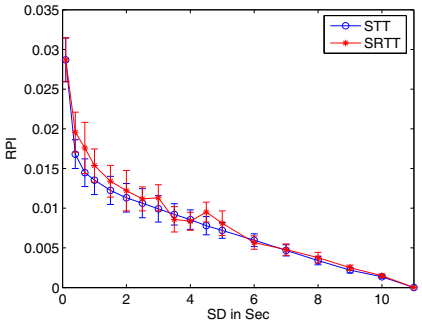

Fig. 3. RPI scores of STT and SRTT schemes vs SD for packet size distribution.

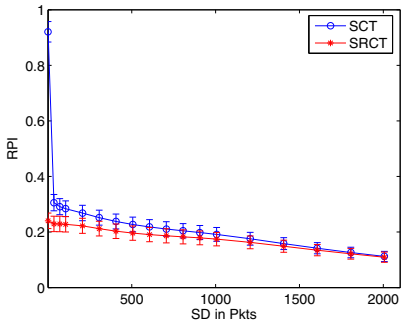

Fig. 4. RPI scores of SCT and SRCT schemes vs SD for packet size distribution.

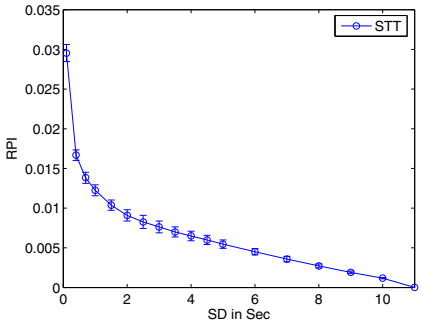

Fig. 5. RPI scores of STT scheme vs SD for inter-arrival time distribution.

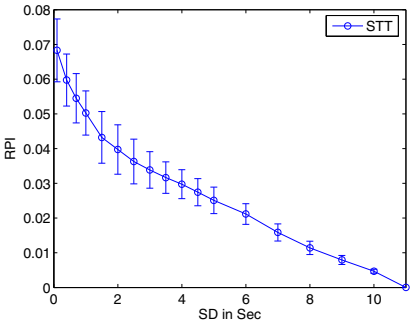

Fig. 6. RPI scores of STT scheme vs SD for packet data rate distribution.

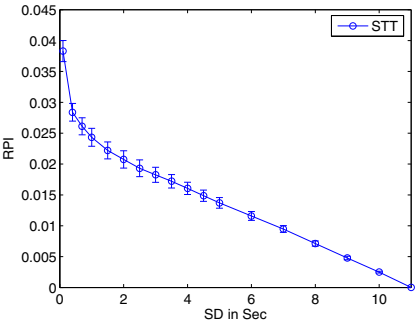

Fig. 7. RPI scores of STT scheme vs SD for packet RSSI distribution.

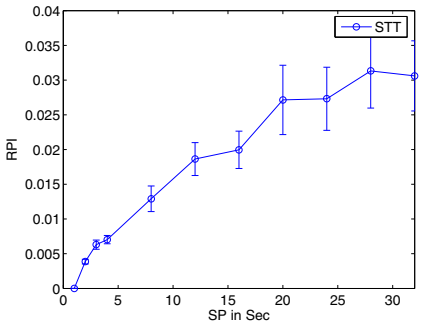

Fig. 8. RPI scores of STT scheme vs SP for packet size distribution.

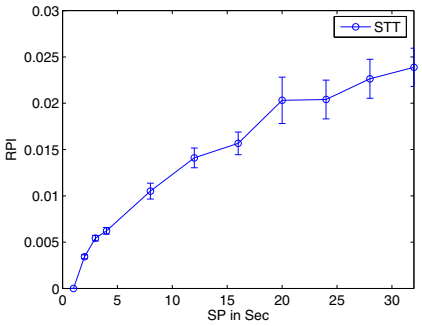

Fig. 9. RPI scores of STT scheme vs SP for inter-arrival time distribution.

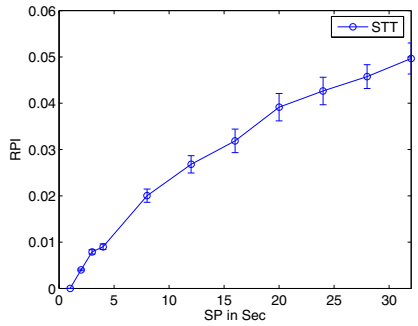

Fig. 10. RPI scores of STT scheme vs SP for packet RSSI distribution.

proportionally to the traffic load as SD is defined as a time interval. From the above results, we conclude that STT is the best sampling strategy in terms of sampling accuracy and easy of implementation in real wireless traffic sensing systems. We now employ STT sampling scheme and compute RPI scores for samples of packet inter-arrival times, packet data rates, and packet RSSI traces. Figures 5, 6, and 7 show their corresponding RPI scores. From these figures (including Figure 3) we can observe that inter-arrival time distribution having the lowest RPI values. If we recall Equation 2, RPI is a function of number of bins (N) and inter-arrival time distribution contains the highest N value. Out of the four traffic metrics studied, samples of packet data rates have higher mean sampling error (5% when SD=1 sec). Since the maximum sampling error is still in acceptable range, we can configure the monitoring nodes to perform STT sampling with SD=1 sec and SP=11 sec for accurately sensing all the traffic metrics in multi-channel wireless network environments.

B. *Effect of Sampling Period*

In this experiment we study the effect of sampling period on the sampling accuracy for various traffic metrics. We keep SD constant at 1 sec and compute RPI scores of samples obtained from STT sampling scheme by varying SP from 1 sec to 32 secs. Figures 8, 9, and 10 show the corresponding RPI scores of traffic metrics. As expected RPI scores increase with increase in SP due to decrease in sample size. We can also observe that packet inter-arrival time distribution is having the lowest RPI values. We also conducted similar experiments by varying channel number for four weeks in our campus. Results are very similar to what we presented in this paper. Due to lack of space, we exclude those results in this paper.

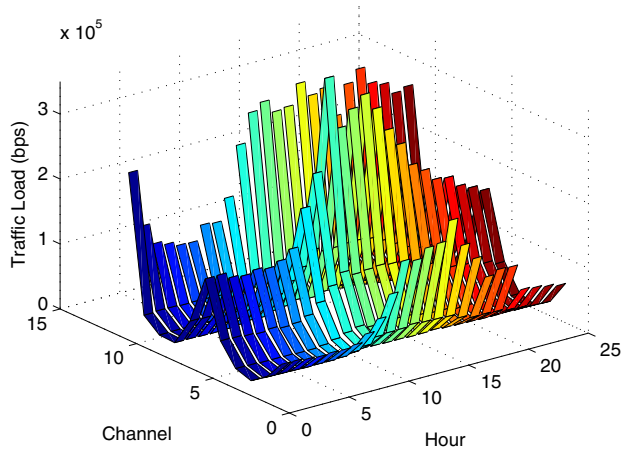

Fig. 11. Working days Traffic load in a Campus environment.

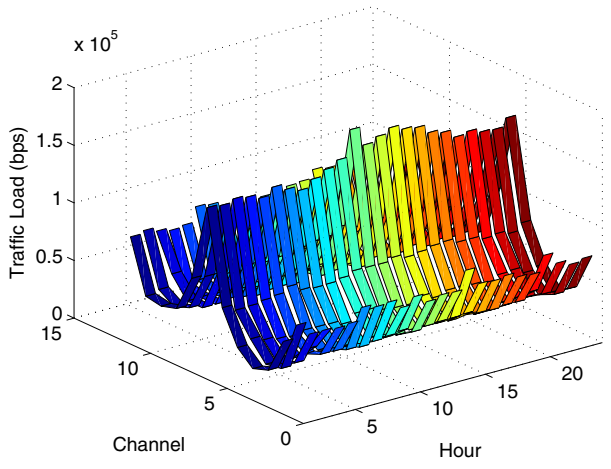

Fig. 12. Holidays Traffic load in a Campus environment.

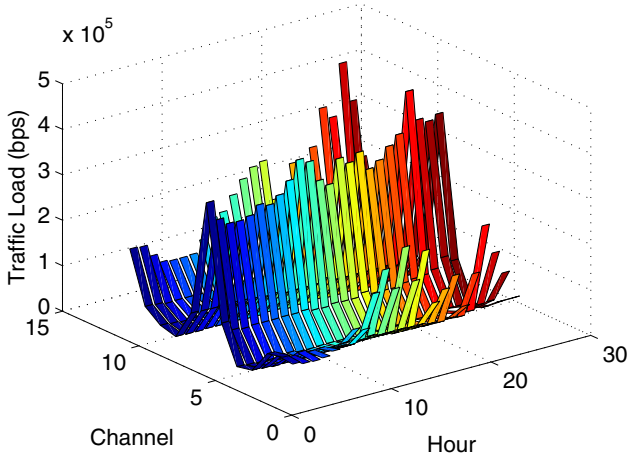

Fig. 13. Working days Traffic load in a Residential environment.

C. *Historical Traffic Statistics*

We implemented STT sampling scheme in all our CalNodes with SD=1 sec and SP=11 sec. One CalNode which is deployed in a residential environment also sends its samples to the CogNet repository via a wired Internet connection. We make use of the sampled traffic to estimate population traffic characteristics like mean traffic load, mean data rate, and traffic intensity. Due to space constraints, here we only show traffic load statistics. Figures 11 and 12 show mean traffic loads (in bps) during normal working days (averaged over 28 working Mondays after excluding holidays and examination periods) and holidays (averaged over 28 sundays) for all 11 channels in UCSD campus. Even though non-overlapping channels (1, 6, and 11) contain higher traffic loads, other channels also face traffic to a certain extent due to leakage of wireless signals into adjacent channels and presence of some experimental testbed APs on them. Further, there is a significant difference in traffic load patterns between working days and holidays. While mean traffic load increases during business hours of working days, it stays pretty much constant for the whole day during holidays. It is to be noted that traffic load is not zero for holidays and non working hours as the APs of UCSD production network always send periodic beacons. Figure 13 presents the working days traffic behavior for all 11 channels in the residential environment. This is obtained by averaging the samples over 28 working Mondays. This underlines the traffic difference between residential environments and campus environments. Similarly, the Figure 14 shows holiday traffic load for all 11 channels in the residential environ-

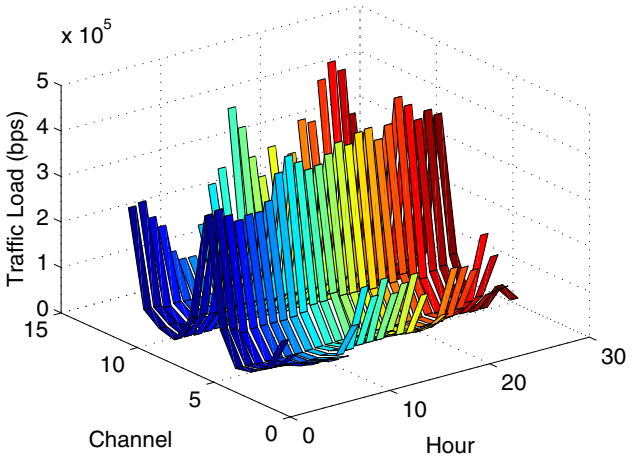

Fig. 14. Holidays Traffic load in a Residential environment.

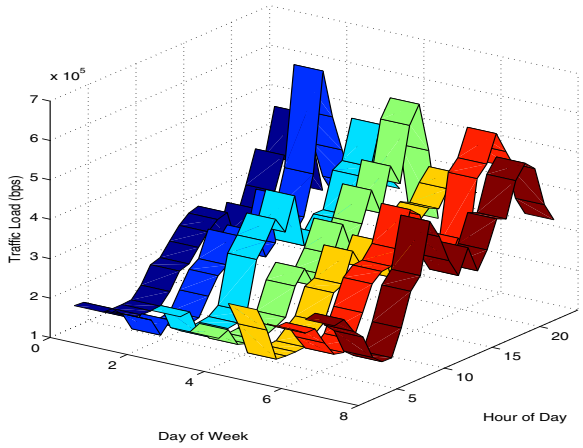

Fig. 15. Channel 1 Traffic load in a Residential environment.

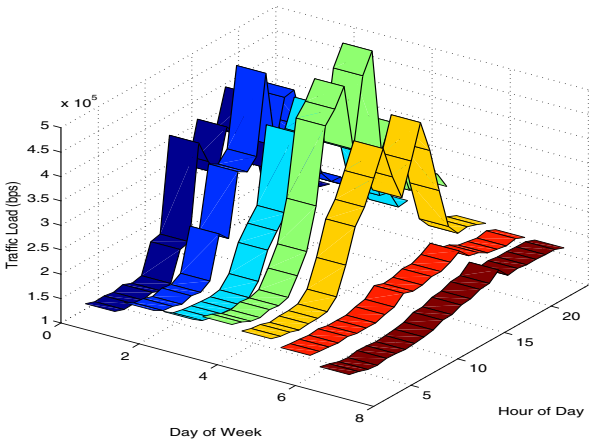

Fig. 16. Channel 1 Traffic load in a Campus environment.

ment, which is significantly different from that shown in Figure 12 for the campus environment. Figures 15 and 16 show mean traffic loads in channel 1 for all 7 days of the week in residential and campus environments, respectively. From these figures we can conclude that there is a significant difference in traffic load patterns between campus and residential environments for all 7 days of the week.

V. Conclusions

In this paper, we proposed a new metric RPI for measuring accuracy of sampling schemes. From our experiments, we found that the Systematic Timer-driven Time-based (STT) sampling scheme is the ideal sampling strategy for traffic sensing and characterization in multi-channel wireless networks. Traffic sensing and characterization is an important and critical building block in the design of Cognitive networking systems. Our future work includes the design of a Cognitive network controller which can observe the surrounding environment (in this case, the wireless medium) to understand the current channel conditions and status of the network, and finally selects the most desirable network configuration.

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] Bheemarjuna R Tamma, B. S. Manoj, and Ramesh R. Rao, "On the Accuracy of Sampling Schemes for Wireless Network Characterization", *in Proc. of IEEE WCNC 2008*, April 2008.

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

[4] 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.

[5] 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.

[6] R. L. Plackett, "Karl Pearson and the Chi-Squared Test", *International Statistical Review*, vol. 51, No. 1, pp. 59-72, April 1983.

[7] J. Fleiss, "Statistical Methods for Rates and Proportions", John Wiley, 2nd Edition, pp. 38-46, 1981.

[8] K. T. Chuang, K. P. Lin, and M. S. Chen, "Quality-Aware Sampling and Its Applications in Incremental Data Mining", *IEEE Transcations on Knowledge and Data Engineering*, vol. 19, no. 4, pp. 468-484, April 2007.

[9] *http://calnode.calit2.net*