Analysis of WLAN Traffic in the Wild

Caleb Phillips and Suresh Singh

Portland State University, 1900 SW 4th Avenue
Portland, Oregon, 97201, USA
{calebp,singh}@cs.pdx.edu

Abstract. In this paper, we analyze traffic seen at public WLANs "in the wild" where we do not have access to any of the backend infrastructure. We study six such traces collected around Portland, Oregon and conduct an analysis of fine time scale (second or fraction of a second) packet, flow, and error characteristics of these networks.

Keywords: Measurement, WLAN, passive monitoring, traffic modeling.

1 Introduction

Analysis of the MAC-level behavior of WLANs is required in order to better deploy and design future systems. To this end, collection and analysis of traffic traces is an important task. The main research reported in this paper analyzes traffic traces collected using a commercial sniffer VWave [1] which has nanosecond time resolution. We characterize the packet level and flow level behavior of these traces and note significant similarities. This result is good news, in that the statistical models we derive can be widely applied for simulations. Our work differs from prior work which have considered congested WLANs at conferences and long-term, coarse-resolution, datasets [2] in favor of studying lightly loaded public hotspots at high resolution, which we conjecture are the norm and not an exception.[1]

2 Data Collection Methodology

The Veriwave WT20 hardware [1] consists of two 802.11 reference radios, real-time linux, and two processors. The WT20 provides nanosecond resolution and it logs the time when it began seeing a frame and the time when the frame finished arriving.

We face two challenges in data collection: The first is placement of the VWave sniffer. Because it has a lower effective receiver sensitivity than most access points today (-75dBm versus -90dBm), we must prevent a large possible packet loss with careful antenna choice and placement. The second problem is practical – we had to obtain permission from the three merchants and further needed to

---

[1] This work was funded by the NSF under grant no. 0325014.

Table 1. Gross statistics of the captures

Capture Name	Length (hours)	Total Pkts	Range	802.11 Mgt.	IP	TCP	UDP	Users Mean	Max
psu-cs (at PSU)	1	127901 (35 pps)	6-771 pps 7k-7Mbps	41543	73473	8803	5965	2.6	5
library (at PSU)	4	696811 (48 pps)	5-672 pps 4k-7Mbps	159699	297481	190962	105405	2.1	3
cafeteria (at PSU)	4	1431897 (99 pps)	7-1318 8k-10Mbps	169541	1026304	911549	108474	10.2	19
pioneer-sq (outdoor)	4	307880 (21 pps)	1-265 1k-3Mbps	206526	99011	94066	4734	2.5	4
urban-grind (coffee shop)	2	490528 (58 pps)	10-355 6k-3Mbps	87423	390514	350034	38696	6.9	9
powell (book store)	4	762574 (53 pps)	8-296 6k-2Mbps	150622	565689	529228	20345	3.4	7

ensure that our equipment was as unobtrusive as possible so as not to affect the "normal" behavior of the users. We collected data at six different locations of which three were located on-campus and three off-campus. Table 1 lists the traces with some gross statistics.

3   Detailed Data Analysis

Our analysis is organized into four categories: network load in terms of users and their residing times, anaysis of MAC-layer errors, the packet arrival process, and finally flow arrival processes and duration times.

3.1   User Load

We consider the number of users over time and the average time spent by a user in the WLAN. We identify the presence of users by a successful DHCP ACK. User departures are indicated by the last message seen with a given MAC address. Table 1 gives the mean and maximum number of users for each capture.

The second statistic we consider is the length of time users stay active in the WLAN (see Table 2). Residing time in four cases fits an exponential distribution ("exp") and in two cases fits the weibull distribution ("wbl"). The quality of the fits is very good as measured by the *deviation* metric Λ [3][2]. Indeed, for all the fits reported, $\Lambda < 0.25$.

Table 2. Residing time of users

	psu-cs	library	cafeteria	pioneer-sq	urban	powell
mean	1363s	5057	3675	4124	4896	2471
max	3486s	13001	11878	12911	8342	8081
std	1203	4900	3332	4744	2969	2181
fit	exp	exp	exp	wbl,$[a, b] = [3276, 0.68]$	wbl,$[a.b] = [5283, 1.44]$	exp

---

[2] This metric for determining the quality of a fit for traffic analysis was first used in [3] where the author explains the rationale behind using this metric rather than a chi-square metric or other metrics.

3.2 Error Analysis

The lower receiver sensitivity of the VWave, makes FCS a poor choice for error analysis. Instead, for the remainder of this section we use MAC retransmissions as an indicator of error and not the FCS value (this is consistant with [4]). We observe a moderate linear correlation between MAC retransmissions and load, with correlation coefficients of 0.53 and 0.54 for packets/sec and bytes/sec respectively. We do see a general reduction in MAC retransmissions with improving rssi, but the relationship does not have a clear fit. And hence, neither load nor signal strength can be conclusively named as a cause of error.
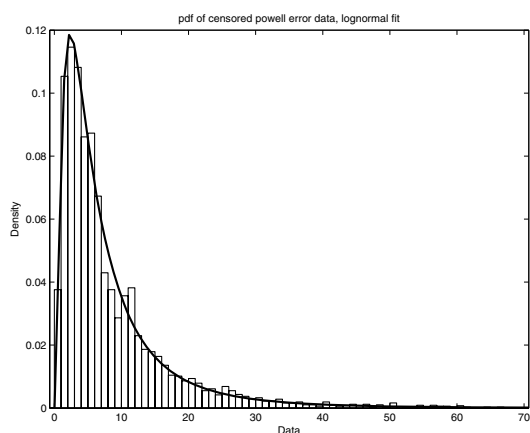

Fig. 1. Censored error data PDF and fit

To fit a distribution to the error process, we compute the probability of error/sec (i.e., for each second what is the probability that a packet will be in error) and then use this set of data to fit a distribution. We note that the probability of no error is quite high and thus any distribution fitting will fail. We therefore resort to a simple technique where we censor the data. To explain this, consider Figure 1 which shows the PDF and a lognormal fit of censored error data. The error data is censored as follows: we have 14400 seconds of data of which 7700

Table 3. Fit for censored error data. The second column represents probability of zero retransmissions/second. The third column is the fit for the data when there is a non-zero probability of retransmission. For the fit parameters we use standard notation.

	Censored data	Fit for censored data
psu-cs	0.63	lognormal $[\mu, \sigma] = [1.8, 0.75], \Lambda = 0.09$
library	0.66	lognormal $[\mu, \sigma] = [1.87, 0.75], \Lambda = 0.15$
cafeteria	0.31	lognormal $[\mu, \sigma] = [2.0, 0.9], \Lambda = 0.07$
pioneer	0.32	gamma $[a, b] = [3.5, 5.3], \Lambda = 0.45$
urban	0.65	lognormal $[\mu, \sigma] = [1.25, 0.85], \Lambda = 0.21$
powell	0.5	lognormal $[\mu, \sigma] = [1.71, 0.93], \Lambda = 0.09$

seconds saw no MAC layer retransmissions (50%). The PDF shown corresponds only to the times when there were MAC layer retransmissions. Table 3 summarizes the fit observed for all six traces after censoring. *It is interesting to note that in all cases except one, the best fit for the censored data is a lognormal fit with parameters $[\mu, \sigma]$ that are relatively close.* Indeed the fits are very good as indicated by the deviation metric Λ. The one exceptional trace, pioneer, is our only capture of an outdoor node – this may serve to explain the different error process observed.

3.3   Packet Arrival Analysis

The metric we consider here is the number of packets/second seen in each trace (the bytes/sec metric follows the same distribution in all six cases). Table 4 summarizes the best distributional fit for each of the six traces. We see that for half the traces t-location scale gives a good fit and for the other half inverse gaussian provides a good fit. Interestingly, the three traces following the inverse gaussian fit correspond to a cafeteria – one at the university, one at a bookstore and a third which is a coffee shop. The three traces that follow t-location scale were generally characterized by few average users (2.1 – 2.6) and lower packet rates which caused non-stationarity.

Table 4. Distribution fits for pkts/sec

	Mean pkts/sec	Fit	Fit parameters	Deviation (quality)
psu-cs	35.3	t-loc scale	$[\mu, \sigma, \nu] = [20, 5.16, 1.09]$	$\Lambda = 2.2$
library	48.3	t-loc scale	$[\mu, \sigma, \nu] = [32.6, 6.1, 1.1]$	$\Lambda = 0.62$
cafeteria	99.3	inv gaussian	$[\mu, \lambda] = [99.3, 75.9]$	$\Lambda = 0.28$
pioneer-sq	21.3	t-loc scale	$[\mu, \sigma, \nu] = [14.1, 3.3, 1.1]$	$\Lambda = 0.79$
urban-grind	58.1	inv gaussian	$[\mu, \lambda] = [58.1, 97.3]$	$\Lambda = 0.33$
powell	52.9	inv gaussian	$[\mu, \lambda] = [52.9, 36.7]$	$\Lambda = 1.4$

3.4   Flow Analysis

Flows are more representative of user behavior than are packet traces, and thus, it is important to consider various flow metrics as well when comparing different traces. We use two flow metrics in this study – flow arrival rate (number of flows/sec) and flow duration (seconds). We do not consider flow interarrival time distribution because the flow arrival rate metric is a cumulative metric based on the flow interarrival times.

To determine flows, we proceeded as follows: we consider pairs of communicating IP address/port tuples and then identify as flows sequential packet exchanges where there were no gaps greater than $t = 64$s. Flow duration is computed based on a time difference between the first and last packet seen. Table 5 summarizes the distribution fit for flows/sec and flow duration. Four traces exhibit the same negative binomial distribution for flow duration. The exceptions are the library and cafeteria traces. The flow arrival rate distributions, on the other hand, show much more variation.

Table 5. Flow distribution fit

	Flow arrival rate	*Parameters*	*Flow duration*	*Parameters*
psu-cs	exponential	$\mu = 11.18$ $\varLambda = 2.9$	neg binomial	$[r, p] = [19.1, 0.69]$ $\varLambda = 0.26$
library	t-loc scale	$[\mu, \sigma, \nu] = [6, 4.6, 0.7]$ $\varLambda = 1.1$	inv gaussian	$[\mu, \lambda] = [9.5, 1.9]$ $\varLambda = 2.3$
cafeteria	exponential	$\mu = 9.6$ (flows/100s) $\varLambda = 0.8$	weibull	$[a, b] = [6.7, 0.78]$ $\varLambda = 6.5$
pioneer-sq	generalized extreme value	$[k, \sigma, \mu] = [2, 4.3, 1.9]$ $\varLambda = 2$	neg binomial	$[r, p] = [0.56, 0.06]$ $\varLambda = 0.9$
urban-grind	neg-binomial	$[r, p] = [0.1, 0.005]$ $\varLambda = 0.18$	neg-binomial	$[r, p] = [0.58, 0.35]$ $\varLambda = 2.5$
powell	neg binomial	$[r, p] = [0.018, 0.007]$ $\varLambda = 2.9$	neg binomial	$[r, p] = [0.7, 0.03]$ $\varLambda = 11$

Our results for the flow arrival process contrast sharply with the results of [5] where the authors find that a weibull distribution fits the observed data. However, their result was based on an hourly scale (i.e., number of flow arrivals/hour) whereas our results model flow arrivals/sec. Our results can thus be used for fine time-grained modeling while their results can be used at larger time scales (hours, days).

A second result from [5] shows that flow duration as measured by number of packets in the flow is lognormal. We measure flow duration by time and in our case we generally see a negative binomial distribution with two exceptions. One possible reason for the difference in results is the definition of flows. Unlike [5], we split a TCP flow into more flows if there is a lull in packets exceeding 64s. In other words, idle times ("thinking") may result in separate flows for the same TCP connection. This model of defining flows has previously been used in Internet traffic modeling [6].

4   Conclusion

The broad results of our analysis are as follows: user residing times can be well modeled by an exponential distribution, packet errors generally follow a lognormal distribution (censored data), load in packets/sec can be modeled using an inverse gaussian distribution (though for very lightly loaded networks t-location scale provides a better fit), flow duration are mostly negative binomial while flow rates do not follow a common distribution. We can conclude that despite the diversity of the WLANs monitored, the users generally are similarly behaved, which is a very useful result from the point of view of future analysis.

References

1. VeriWave: http://www.veriwave.com (February 13 2007)
2. CRAWDAD: http://crawdad.cs.dartmouth.edu (February 13 2007)
3. Paxson, V.: Empirically derived analytic models of wide-area tcp connections. IEE/ACM Transactions on Networking **2**(4) (August 1994) 316 – 336

4. Rodrig, M., Reis, C., Mahajan, R., Wetherall, D., Zahorian, J.: Measurement-based characterization of 802.11 in a hotspot setting. In: Proceedings of the ACM SIGCOMM 2005 Workshop on experimental approaches to wireless network design and analysis (E-WIND-05). (2005)
5. Meng, X., Wong, S., Yuan, Y., Lu, S.: Characterizing flows in large wireless data networks. In: ACM MOBICOM. (Sept 26 – Oct 1 2004)
6. Zhang, Z.L., Ribeiro, V., Moon, S., Diot, C.: Small-time scaling behaviors of internet backbone traffic: an empirical study. In: IEEE INFOCOM. (2003) 1826 – 1836