Characterizing Flows in Large Wireless Data Networks

Xiaoqiao (George) Meng*, Starsky H.Y. Wong*, Yuan Yuan‡, Songwu Lu*

*Computer Science Department
University of California
Los Angeles, CA 90095
{xqmeng,hywong1,slu}@cs.ucla.edu

‡Computer Science Department
University of Maryland
College Park, MD 20742
yuanyuan@cs.umd.edu

ABSTRACT

Several studies have recently been performed on wireless university campus networks, corporate and public networks. Yet little is known about the flow-level characterization in such networks. In this paper, we statistically characterize both static flows and roaming flows in a large campus wireless network using a recently-collected trace. For static flows, we take a two-tier approach to characterizing the flow arrivals, which results a Weibull regression model. We further discover that the static flow arrivals in spatial proximity show strong similarity. As for roaming flows, they can also be well characterized statistically. We explain the results by user behaviors and application demands, and further cross-validate the modeling results by three other traces. Finally, we use two examples to illustrate how to apply our models for performance evaluation in the wireless context.

Categories and Subject Descriptors

C.2.3 [**Computer-Communication Networks**]: Network Operations—*Network management, Network monitoring* ; C.2.5 [**Computer-Communication Networks**]: Local and Wide-Area Networks

General Terms

Measurement

Keywords

Network analysis, Flow modeling, 802.11, LAN

1. INTRODUCTION

Wireless data networks based on IEEE 802.11 technology have become increasingly popular in enterprise and campus environments. While there have been several studies on network usage and mobility patterns of wireless networks at the host level [2, 9, 15, 21], the problem of modeling wireless data traffic has remained largely unaddressed. In this paper, we use extensive traces to model wireless traffic at the flow/connection-level The results statistically characterize both static and roaming flows, and are cross-validated using four traces independently collected from different scenarios.

Flow-level modeling has become popular in the wired Internet in recent years [6, 12, 17, 20, 23]. It helps to better understand Internet traffic, and facilitates effective congestion management and traffic engineering. In wireless and mobile networks, extensive studies have shown that flow models are critical to evaluating and refining many networking protocols, such as wireless packet scheduling [24], admission control [25], and wireless TCP [5]. However, most studies in the wireless literature are based on unrealistic settings of static flow configuration or simplistic Poisson model. In fact, as we will show later in this paper, such simulation and analysis models can produce misleading results compared with using the models derived from real traces. In some cases like packet scheduling, the discrepancy can be as large as 99.4%. In other cases like wireless TCP, the performance gain can be over-estimated by an-order-of-magnitude large.

Modeling data flows in wireless networks has to address several issues. These include: highly time-varying traffic due to reduced multiplexing of smaller flow population, location-dependent flow characteristics at different access points (APs), and roaming flows from mobile hosts. Fundamentally, compared with wired networks, wireless flows exhibit rich dynamics in both time and spatial domains, and modeling has to address both aspects.

In this paper, we focus on flow characterization. The main trace used is quite comprehensive. It has not been analyzed before, and consists of 80.8GB data records from 2.17 million TCP flows on 1706 hosts, accessing 476 APs deployed in 161 campus buildings, over a three-month period in 2002 [9]. The other three traces used [2, 15, 21] are independently collected from campus department, corporate and conference scenarios.

We make three contributions in this paper:

- We characterize both static and roaming flows. For static flows, we take a *two-tier* modeling approach to deriving a heavy-tailed, Weibull regression model that accurately approximates the flow arrivals at individual APs. For roaming flows, 97.7% of them have less than five handoffs, and the number of handoffs can be well modeled by the Geometric distribution. For flows at different APs in geographic proximity, they exhibit spatial similarity in arrival distributions.

- We explain the modeling results from the perspective of user behaviors and application demands. For example, the Weibull regression model attributes to both

MobiCom'04, Sept. 26-Oct. 1, 2004, Philadelphia, Pennsylvania, USA.

observations of strong 24-hour periodicity and diurnal cycle in user activity, and coexistence of applications with short- and long-interarrival times. The models are also partially verified using three other traces.

- We use two examples of scheduling and wireless TCP to showcase how to apply our results to evaluate the *dynamic* behavior of network protocols. In the wireless TCP example, the 55% throughput gain in static flow settings diminishes to 7.8% when using our dynamic flow models.

The rest of the paper is organized as follows. In Section 2, we explain the motivation and challenges for flow-level modeling. We also survey the related work. In Section 3, we describe the analysis methodology used in this work. In Section 4 and 5, we characterize static flows and roaming flows respectively. In Section 6, we use Proportional Fair Scheduling and TCP performance as showcases to exemplify the application of our results. We discuss related issues In Section 7. Section 8 concludes the paper.

2. BACKGROUND AND RELATED WORK

In this section, we first explain why we take the flow-level modeling approach rather than packet-level modeling. We then identify three challenges for flow-level modeling and discuss the related work.

2.1 Flow-Level versus Packet-Level Modeling

Two popular approaches to traffic modeling are packet-level characterization and flow-level characterization. In addition to the claimed merits in the wireline context [23, 6], flow-level modeling has the following advantages in wireless networks.

First, packet-level modeling cannot avoid the complex interactions among TCP, IEEE 802.11 MAC with retransmissions, dynamic wireless channel, and applications. In contrast, flow modeling captures the dynamic user demand and application behaviors. Second, if we model the aggregate effect of such interactions at the packet-level, we may not be enable to correctly evaluate new designs like MAC or wireless TCP, since their impact is already unnecessarily reflected by the derived models. In contrast, flow models, together with other complementary models such as wireless channels (e.g., [4]) and network protocols, allow us to evaluate new networking designs in comprehensive wireless settings. Finally, packet-level models cannot inform us of the performance bottleneck of the networking system. In the worst case, the bottleneck may not come from wireless but from the wired portion of the connection. For the above reasons, we take the flow-level approach in this work.

2.2 Challenges for Flow Characterization

Characterizing flows in wireless networks has to address three nontrivial issues. (1) Wireless traffic is highly nonstationary. Although nonstationarity has been conceptually recognized before in wired networks, not much effort has been made to model it since the nonstationarity is alleviated to some extent by the large amount of superposition in the order of thousands of flows or even more [18]. In the wireless domain, flow population is an order of magnitude smaller, which makes the nonstationarity feature prominent. (2) Wireless traffic is location dependent. In wired networks, users in the same subnet roughly share the same view on the network. In wireless, users access the network through geographically-spread APs over the wireless channel, and APs are the main networking component to manage traffic flows. Each AP may perceive different flow characterizations. It is well known that many WLANs have one or two hot-spot APs that receive the largest amount of traffic [21]. Roaming users may also move around different APs, and alter flow characteristics in all the transit APs. (3) User mobility may also incur flow mobility. Roaming hosts cause continuous flows running on them to be roaming too. This brings up the issue of flow handoff.

2.3 Related Work

There has been much work on Internet traffic modeling ([28, 29] the references therein for a sampling of the literature). We only compare with two most relevant to ours. Paxson [28] studied wide-area TCP connections. They proposed to use several heavy-tailed distribution models to characterize the statistical process associated with TCP flows. Feldmann [3] modeled wired TCP flow arrivals by using the 2-parameter Weibull model and claimed that it gives a better fit than other distribution models. Compared with their results, We study wireless traffic and our proposed Weibull regression model captures the time-varying flow traffic in wireless networks.

A recently popular approach to Internet traffic study is to model the traffic at the flow level [6, 23, 27]. However, all such studies use idealized models, e.g., Poisson process, to characterize flows. While such simplified models may be fine in the wired Internet because of large number of multiplexing flows, they are not appropriate for wireless networks, where the traffic is highly nonstationary and flow number is much smaller.

Few studies have focused on wireless traffic modeling. Several recent papers have carefully characterized user and mobility patterns in wireless data networks [2, 9, 10, 11, 13, 14, 15, 21]. However, most of them focused on host-level rather than flow-level. None of them uses multiple traces for cross validation. Specifically, Tang and Baker [14] examined the user and device patterns in a metropolitan-area wireless network. In their another work [15], they studied network activity and host mobility patterns at the Stanford Computer Science Department. Kotz and Essien [9, 10] examined wireless user activity, aggregate traffic and AP/building activity for the Dartmouth campus wireless network. The traces used in [9, 10] were collected in 2001. In a more recent work [11], Henderson, Kotz and Abyzov further collected traces from the same network and compared the network usage with their previous study. Balazinska and Castro [21] studied user population characteristics, load distribution and network usage in corporate networks. Balachandran et al [2] characterized the aggregate network load, aggregate utilization and user patterns in a three-day conference setting. In another work, Chinchilla et al [16] recorded HTTP queries from the campus wireless infrastructure at UNC, and found that the information accessed by HTTP queries exhibits spatial locality. The key difference between these previous studies and ours is that we characterize flows in stead of user behavior or network performance. Our cross-validation using three different datasets, despite incompleteness, is a small step forward toward building generally applicable flow models for wireless data networks.

Dataset	Network type	Network size	Type of traces	Starting time	Duration	Traffic volume
Dartmouth02	Campus-wide	161 buildings, 476 APs	tcpdump, syslog,snmp	02:00am, 28Mar2002	3 months	2.17 million TCP flows
UCSD01	Conference	1 subnet, 4 APs	tcpdump, snmp	08:00, Aug2001	3 days	299k TCP flows
IBM02	Corporate buildings	subnet # unknown, 177 APs	snmp	00:00, 20July2002	4 weeks	1366 hosts
Stanford00	Departmental buildings	1 subnet, 12 APs	own format	20Sep1999	12 weeks	74 hosts 78 million packets

Table 1: Summary of data traces

3. METHODOLOGY

In this section, we first describe the four data traces used and the data analysis techniques applied. We then present a systematic method to model wireless flows.

3.1 Data Traces

Compared with earlier work that each used a single trace, we use four traces together in this work for study and validation. The four wireless traces were independently collected from Dartmouth College [9], SIGCOMM'01 conference [2], IBM research buildings [21] and CS department of Stanford University [15]. Among them, the Dartmouth trace, abbreviated as Dartmouth02, offers the most comprehensive data collection over a large campus-wide network consisting of 476 APs across multiple subnets. Therefore, we use it as the main trace to model wireless flows. The other three traces, i.e., UCSD01, IBM02 and Stanford00, were collected from smaller settings, and some of them do not have complete information regarding the flow and roaming statistics. We use them for validation purpose. A summary of these four traces is given in Table 1.

The Dartmouth02 trace used in this work was collected in Spring 2002. It logs data of *tcpdump, syslog* and *snmp*. The tcpdump log is about 71.9 GB data, containing headers of all transmitted packets from five wireless subnets, i.e., *Berry, Brown, Collis, Sudi* and *Whittemore*. By analyzing syslog traces, we found that these tcpdump logs mainly involve 31 APs in the five subnets: B1, B2, . . ., B13 in Berry; BN1 and BN2 in Brown; C1, C2 and C3 in Collis; S1, S2, . . ., S5 in Sudi; W1, . . ., W8 in Whittemore. 15 out of the 31 APs have more than 30K TCP flows during the three-month period and they are most frequently used in our study.

The syslog data is about 534MB and records the activity of wireless cards, including association, roaming, etc. It logs an accurate event history for each card. The snmp data is about 8.4GB and logs the list of wireless cards associated with each AP during every 5-minute polling period of an AP. It offers an alternative way to track wireless hosts. In our usage of Dartmouth02 dataset, we use syslog data instead of snmp since the former is of finer time granularity.

3.2 Analysis Techniques

Synthesizing traces The first new analysis technique we apply is to synthesize and correlate datasets of tcpdump and syslog. In contrast, previous studies [2, 9, 15, 21] typically analyzed them separately. Using synthesized data allows us to track both the temporal evolution and the spatial location of a flow. Specifically, for each TCP flow recorded in tcpdump, we retrieve the corresponding wireless MAC address, then look up the syslog trace to extract all the recorded syslog messages for the MAC address. By analyzing the syslog messages, we can infer the set of APs that the flow used during its duration. If only one AP is found to be used by the flow, we categorize the flow as a *static flow*; otherwise, it is a *roaming flow*.

Statistical techniques In this paper, we match traces against seven popular distribution models to approximate the underlying statistics. The Probability Density Function (PDF) and Complementary Cumulative Distribution Function (CCDF) of these models are shown in Table 7 of Appendix 11.1. These seven models range from the well-known Exponential and memoryless Geometric distributions, to the heavy-tailed Pareto and Weibull distributions. While by no means do they represent a complete list of distributions, our analysis shows that some of them match fairly well with the traces, thus accurately capturing the underlying statistics. When using these models to approximate the empirical data, we always apply Maximum Likelihood Estimator (MLE) for parameter estimation.

To gauge how well a given analytical model matches the real data, we use the metric *average deviation*, proposed by Paxson in [28]. This metric reflects the fraction of samples that deviate from the analytical model; the smaller the value, the better the model. The calculation of average deviation requires to place the observation data into a number of bins. The method we use to choose bin count and space the bins is the same as that of the statistics software Dataplot [7]. In the analysis, we also use the quantile-quantile (q-q) plot [19] to statistically determine whether two sets of samples follow the same distribution.

3.3 Toward a Systematic Approach to Flow Modeling

The goal of this work is to characterize packet flows. Specifically, we focus on TCP flows, since they dominate all traces, e.g., they contribute 88% of total transmitted bytes in the Dartmouth02 dataset. We seek to address the following three issues: (1) At which level to model flows, the APs or the subnet? (2) how to characterize packet flows? (3) what metrics to measure?

In order to model flows in wireless networks, we decide to model the aggregate flows at each AP rather than for the entire subnet. This choice is made because flow modeling at each AP offers several benefits. First, it helps to better manage radio channel resources, which can be best done at an AP. Second, it helps us to understand the impact of layer-2 handoffs and to manage hot-spots. Finally, AP-level modeling best captures the location dependent feature of wireless traffic, since wireless connectivity for mobile users is provided by AP.

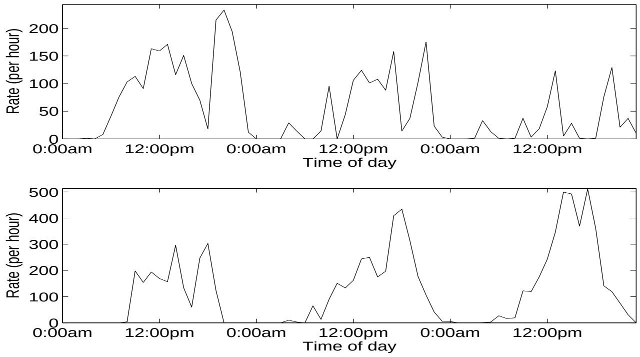


Figure 1: Flow arrival rate for AP W3 and AP W5 between 0:00am April 13, 2002 and 23:00pm April 15, 2002 (EST)

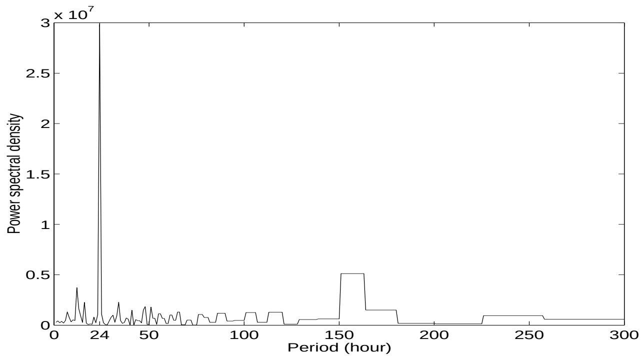


Figure 2: Frequency representation of TCP arrival rate for AP W6 (arrival rate is measured for every 15 minutes)

Our modeling method categorizes flows into static and roaming flows. Modeling static flows is further split into two phases: we first examine the temporal dynamics flows at a given AP, and then study flow characteristics across multiple APs.

Static flows at a given AP are determined by two parameters, flow interarrival time that characterizes the frequency of new flows generated by users, and flow duration that determines its lifetime. It is easy to see that flow duration depends on the average transmission rate perceived by the flow, which further depends on several coupled factors such as channel conditions, behaviors of the concurrent sharing flows, and TCP protocol behavior (i.e., its congestion control actions). Therefore, to simplify the modeling process, we use the data size transferred by the flow to characterize the flow duration; In general, the larger the flow size, the longer the duration. Our analysis in later sections characterizes both metrics.

As for the location dependent features of flows, we study the similarity of flow statistics across multiple APs. In addition, each roaming flow is modeled by two metrics, the number of handoffs during its lifetime, and the residing time of a flow at each transit AP.

4. STATIC FLOWS

This section characterizes static flows. We first examine the temporal evolution of flows at an AP, and then study the spatial correlation of flows across APs. This section answers three questions:

- How do new flows arrive at an individual AP?
- What is the relation between flow arrivals at different APs?
- For the new arrival flows, what is the distribution for their flow duration?

The following three subsections answer these questions. We further verify such results using other traces.

4.1 Flow Arrivals at An AP

4.1.1 Nonstationarity in Coarse Time Granularity

We first show that flow arrivals are nonstationary over coarse-time scales, say, from few hours to a day. Figure 1 provides a snapshot of flow arrivals at two APs. The figure shows that the arrival process varies over days and hours within a single day. More specifically, within each day, the flow arrival rate starts to reach a high value in the morning and drop to nearly zero in the late evening. We observed the same phenomenon with all APs.

The nonstationarity characteristic can be explained by two factors. First, the time-varying flow arrival pattern reflects a typical user's behavior of a diurnal cycle during working hours and off hours within a day. Second, compared with the wired network scenario, the level of multiplexing is significantly lower. In our study, the flow arrival rate ranges from 0 to 600 flows per hour, much smaller than the wired case [29]. Such a reduced level of flow multiplexing exaggerates the nonstationarity.

4.1.2 Two-tier Modeling to Nonstationary Flows

The flow nonstationarity poses challenges for statistical modeling, and renders many results for the wired traffic flows inapplicable. In this work, we take a novel two-tier modeling approach. More specifically, we partition the entire time into coarse and fine scales. At the fine scale (say, within an hour), we found that the trace can be well approximated by simple statistical distributions. At the coarse scale (say, several hours to days), we use regression model of the statistical distribution to capture the time-varying feature.

The above two-tier modeling method is made possible by the observation that the flow arrival process, despite nonstationarity, exhibits a 24-hour periodic mode. To illustrate how strong this periodicity is, we plot the power spectral density (PSD)[1] for the arrivals at AP W6 in Figure 2, The figure shows a dominant peak, valued at 24-hour period, in the frequency domain. The ratio between the peak and the average PSD value is as high as 17.7. We have also examined other 14 APs and found that, the peak is still reached at 24-hour cycle and the ratio between the peak and the average PSD is between 13.7 and 40.0. This frequency-domain calculation clearly demonstrates the dominant mode at the 24-hour period.

[1] Since the trace contains only 77 days' data, we only plot the PSD values in higher frequency range $[\frac{1}{300hr}, +\infty)$.

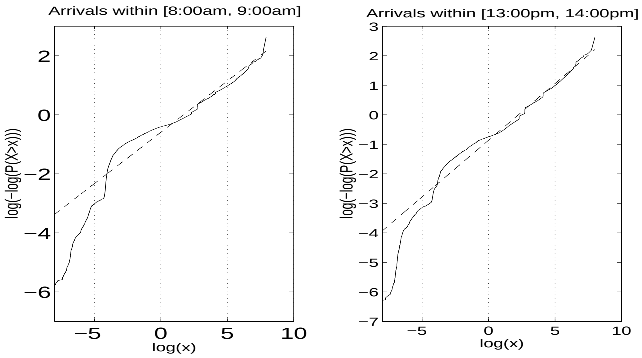


Figure 3: $\log(-\log(P(X > x)))$ **vs** $\log(x)$ **for AP W6 (**x **is flow interarrival time)**

We further notice that the second highest PSD is typically achieved around 160-hour period point, roughly corresponding to the weekly periodicity. However, the ratio between the second highest PSD value and the average PSD for all the 15 APs lies between 1.3 and 2.7, much smaller than the ratio for the 24-hour point. We thus ignore such a high-order component in our analysis.

The above dominant 24-hour mode attributes to the user behavior. Typical users will follow a diurnal cycle of working hours and off hours within a day; this leads to the 24-hour periodicity. Note that, we are not claiming this is a unique feature in wireless data networks. In fact, such a phenomenon has been also observed in a wired scenario [29].

The above analysis lays the foundation for our two-tier modeling method. As long as we can characterize the daily 24-hour process, we may arrive at a reasonably good approximation model across days, while ignoring other high-order components.

4.1.3 Weibull Distribution at Fine-time Granularity

Our next finding is that the interarrival times can be well modeled by a Weibull distribution at fine-time scales, e.g., hourly basis. This result holds for all 24 hourly intervals.

To this end, because of the dominant 24-hour mode, we divide a day into 24 hourly intervals and merge all the interarrival time samples within the same hour-of-day. For each hourly interval, we perform a statistical test to see whether the Weibull model can give a good fit. The statistical test is to plot $\log(-\log(P(X > x)))$ versus $\log x$. If the real data follow the Weibull distribution, the plotting should produce a straight line since the CCDF function for Weibull distribution is $P(X > x) = e^{-\mu x^k}$. By using the best-fit linear regression method [19], we find that 86.4% of all samples produce a R^2 value [2] larger than 0.90, which implies a close approximation to lines. For e.g., Figure 3 gives the plotting when applying the test to two hourly intervals, [8:00am, 9:00am] and [13:00pm, 14:00pm], for AP W6. In the study, we also tested other five continuous distribution models: Exponential, Lognormal, Gamma, Pareto and Extreme-value. None of them produces a consistently good match. Our further analysis shows that, if further reduce the granularity into fine scales, say, half an hour, the Weibull model still matches well. However, if we increase the granularity to coarser scales, say, two hours, simple distributions such as Weibull model are generally not sufficient because the nonstationarity makes the traffic much more variable. Therefore, we select the hourly scale in modeling. We further find that our conclusion holds for all APs.

The heavy-tailed, Weibull distribution model can be explained based on application behaviors. It demonstrates the coexistence of both short and long interarrival times, which are commonly observed in data applications. Moreover, it invalidates the memory-less Poisson model, which is still used in some analysis and simulations today (e.g., [8, 24]).

4.1.4 Weibull Regression Model for All Time Scales

We have seen that the Weibull model is appropriate in the fine-time granularity. However, when extended to coarse-time granularity, the Weibull model has a limitation in the sense that its location parameter μ varies significantly. In order to model flow arrivals in *all time scales*, we should incorporate the time-of-day effect on the Weibull parameters in our model. This leads to a Weibull regression model. The model is a direct application of the two-tier modeling method. It leverages the dominant 24-hour periodic mode and the diurnal cycle within a day, as well as Weibull distribution within an hour.

The Weibull regression model can be represented by the following PDF function :

$$f(x|t) = k\mu(t)x^{k-1}e^{-x^k\mu(t)}, \qquad (1)$$

where x denotes the flow interarrival time, t is the time-of-the-day measured in hours, $\mu(t)$ and k are the Weibull parameters. $\mu(t)$ reflects the time-varying arrival rates at different times-of-the-day. Since the arrival process matches the diurnal cycle of human daily activity, we tailor the shape of $\mu(t)$ to capture this feature:

$$\mu(t) = \begin{cases} \mu_0, & t \in [t_h, t_l] \\ \mu_0 - \beta(t - t_l)(t_h + 24 - t), & t \in [0, t_h] \cup [t_l, 24] \end{cases} \qquad (2)$$

where $[0, t_h] \cup [t_l, 24]$, $[t_h, t_l]$ represent the typical high- and low-traffic periods within one day respectively. In the above expression, μ_0 is a fixed value, representing the typical arrival rate during the high-traffic period. $\mu(t)$ remains μ_0 during the high-traffic period $[t_h, t_l]$, expecting that the distribution of interarrival times does not vary much. In the low-traffic periods $[0, t_h] \cup [t_l, 24]$, $\mu(t)$ is gradually decreasing up to μ_0 when the time-of-day is approaching either t_h or t_l, consistent with the observation that the arrival rate is likely to increase when the time-of-day t is approaching morning working hours. Another parameter β in (2), intuitively determines how rapidly the arrival rate decays when departing from the high-traffic period.

The complete Weibull regression model combines both (1) and (2). In total it has 5 unknown parameters: $k, \mu_0, \beta, t_l, t_h$. Details of estimating these 5 parameters are provided in Appendix 11.2.

In the following, we evaluate the accuracy of the proposed Weibull regression model. Our result shows that the model accurately approximates the real traces.

Our evaluation method is to compute the discrepancy between the model and the real data. The time granularity used is still an hourly interval, and the real flow arrivals are

[2] R^2 is the *coefficient of determination*. It is used to measure the adequacy of a regression model.

Subnet	Berry				Brown		Collis		Whittemore						
AP	B4	B9	B11	B12	BN1	BN2	C1	C2	W2	W3	W4	W5	W6	W7	W8
β	0.013	0.028	0.019	0.015	0.006	0.005	0.010	0.011	0.002	0.008	0.007	0.001	0.004	0.005	0.006
μ_0	0.85	0.98	0.95	1.22	1.31	1.48	0.74	0.70	0.09	0.21	0.18	0.08	0.28	0.11	0.38

Table 2: Parameter β and μ_0 for the 15 APs

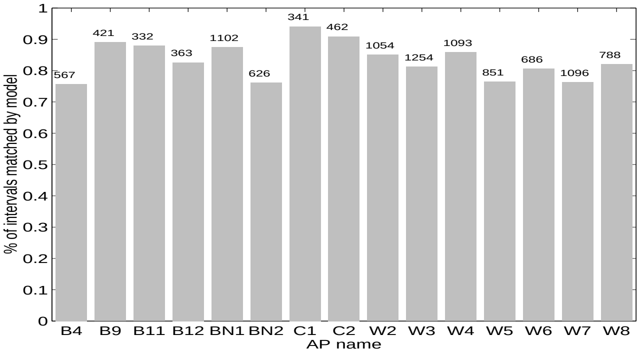


Figure 4: Evaluate the model accuracy for the 15 APs (the value above each bar represents the number of involved intervals for the AP)

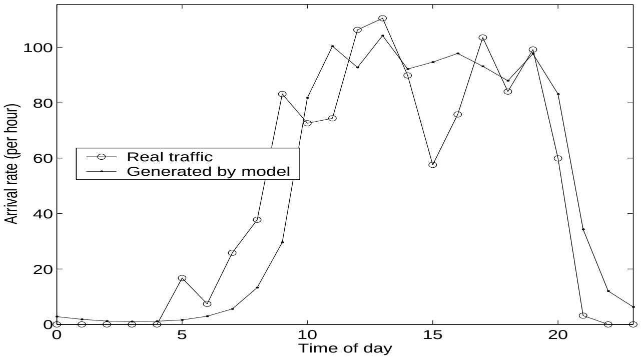


Figure 5: Flow arrivals from the model and the trace

Metric	Type of AP pair	#	Ratio of passing null hypothesis
Interarrival times	Same subnet	74	90%
	Diff subnet	251	36%
TCP data size	Same subnet	74	50%
	Diff subnet	251	57%

Table 3: Spatial similarity for the distribution of flow interarrival times and flow size

placed into an hourly bin for each of the 24-hour period. We apply the quality-of-fit technique based on average deviation (described in Section 3.2) to all APs. The fraction of time intervals having good match with the model (i.e., the average deviation is smaller than 0.5) is plotted in Figure 4. In the figure, the number above each bar denotes the number of time intervals with sufficient observations (≥ 10). We conclude that during 76.1% to 94.0% intervals, flow arrivals well modeled by the proposed Weibull regression model.

The accuracy can also be roughly examined by comparing the sample paths from the real trace and from the model. Figure 5 plots the arrivals generated by the model and the arrivals from the real trace at AP W6. In the figure, we notice a saddle point (at 15:00pm) in the real trace is not well approximated by the model, but in overall, the Weibull regression model captures the important features of the actual flow arrival process, e.g., the sharp switching between the high- and low-traffic periods and the average arrival rate during the high-traffic period.

4.2 Spatial Similarity across APs

After describing models for flow arrivals at a given AP, we further examine spatial correlation of flow arrivals across APs.

Our main finding is that different APs in the same subnet tend to possess identical distribution for flow arrivals. The statistical tool used for identical distribution test is q-q plot. Among the entire 31 APs, 26 APs have more than 8000 flow arrivals in the 3-month period and they are used in the test. Such 26 APs belong to 5 subnets and they form 74 intra-subnet pairs. We apply q-q plot to each of these 74 pairs, to examine whether the two APs have similar distribution of flow arrivals or not. The results are presented in Table 3. The table shows that 90% of the AP pairs within the same subnet are likely to have the same distribution of flow interarrival times since they pass the null hypothesis at the significance level $\alpha = 0.01$.

Since the flow interarrival times at individual APs can be modeled by the Weibull regression model, the above observation also indicates that, intra-subnet APs should have similar parameter values in the model. Values given in Table 2 indeed confirm this conclusion. Essentially, our results show that APs in the same subnet share spatial similarity in the sense that their flow arrivals follow approximately the same statistical distribution.

The observed spatial similarity can be explained by similar network usage patterns across APs. APs in the same wireless subnet are typically deployed in a geographic proximity. Users accessing these APs tend to belong to the same interest group, e.g., students taking the same course or being in the same year class to share a residence hall. These users in the same group tend to exhibit similar usage behaviors. Take the subnet Whittemore as an example, Whittemore is located in a student residential building. In such an environment, students may mainly use the wireless access to check their emails. Therefore, it is not surprising to find that POP3 is the dominant (in terms of number of connections) application type in all APs within Whittemore. Another recent work [16] has found that significant correlation exists in the information categories requested by nearby mobile users. Their finding partially corroborates our analysis.

Table 3 also shows that 36% AP pairs from different sub-

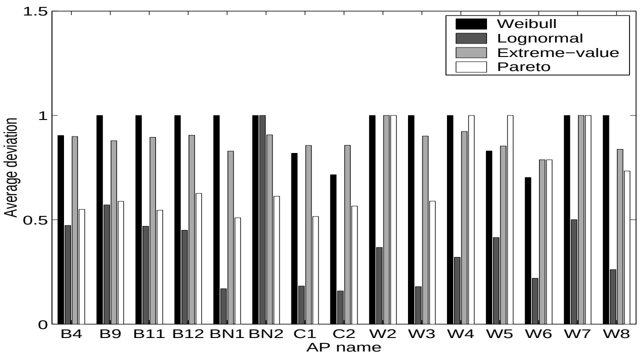


Figure 6: Quality of fit for flow size

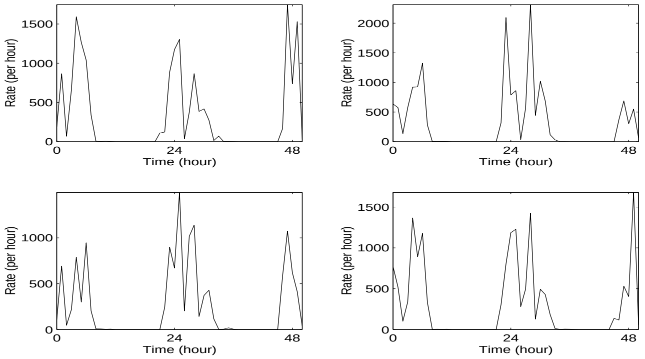


Figure 7: TCP connection arrival rate varies over time (for the 4 APs in UCSD01 dataset)

nets observe similar flow arrival patterns since they can pass the null hypothesis. This indicates that in a campus environment, many users spread across different subnets may also belong to the same interest group and share similar access pattern. Nevertheless, the similarity in this case is not so dramatic as at the subnet level.

4.3 Modeling Flow Duration via Flow Size

As we argued in Section 3, we indirectly characterize flow duration via the flow size, i.e., the total bytes transferred by the flow. We model the statistical distribution of flow size at an AP, and examine the spatial similarity across APs.

It turns out that the flow size can be best approximated by the Lognormal distribution. Figure 6 shows that such a model provides the best fit. The two parameters for the Lognormal distribution, μ and σ, are estimated to be ranging between $[7.37, 8.50]$ and $[0.98, 1.92]$, respectively. The best fit of the Lognormal model indicates that the distribution for flow size is heavily tailed. Intuitively, this distribution means that most flow sizes are reasonably small while a few others are very large.

We then examine whether the distribution of flow size also exhibits spatial similarity. However, we could not find credible evidence. The result of applying q-q plot is given in Table 3. No matter whether the AP pairs are from the same subnet or not, the fraction of AP pairs passing the null hypothesis is roughly the same. One explanation is that, most TCP connections are communicating with hosts (or requesting services) over the Internet. Consequently, the TCP data size is largely affected by the Internet-wide metrics (e.g., file size distribution), rather than the local wireless-domain.

We also make an interesting observation in the analysis. We find that about 97.3% flows remain non-idle during their lifetime. However, 2.7% flows experience idle periods longer than 5 seconds [3] during their lifetime. Moreover, 66% of such idle flows experience the idle period only once in their lifetime. The idle flow phenomenon can be partially explained from the application perspective. Protocols like HTTP/1.1 [1] allows persistent connections to transmit multiple small data objects when accessing the same server. However, the small percentage of idle flows does not seem to carry much weight in flow models.

4.4 Cross-validation Using UCSD01 Trace

We now use the dataset UCSD01 to cross-validate the results observed from the Dartmouth02 trace. The dataset UCSD01 provides both tcpdump and snmp traces for four APs (see Table 1). We use it to validate both the flow arrival and the flow size models.

Flow arrivals at each AP are nonstationary, as shown in Figure 7. However, similar to the Dartmouth02 trace, they also exhibit a diurnal cycle. Furthermore, We find that the Weibull regression model is still valid in this case, but the parameters are different from the Dartmouth network. Figure 8.1 shows the accuracy of the regression model at each AP. For 80% to 99% intervals, the real data match the model with an average deviation less than 0.5. The parameters for the Weibull regression model are $\beta \in [0.456, 1.181]$, $\mu_0 \in [0.34, 0.44]$, varying among the four APs. We can make a comparison between these values and their counterparts for the Dartmouth02 dataset, which is given in Table 2. First, β is much larger than in the Dartmouth02 case, indicating that the arrival rate in the UCSD01 case decades much faster when departing from high-traffic time period. This coincides with the fact that the UCSD01 dataset is from a conference WLAN wherein people rarely use network after the official meeting time is ended. Second, μ_0 in the UCSD01 dataset is larger than the subnet Whittemore in the Dartmouth02 case while smaller than the subnets Berry, Brown, Collis. This shows that the average flow rate in a conference WLAN is comparable to WLANs within a campus-wide network.

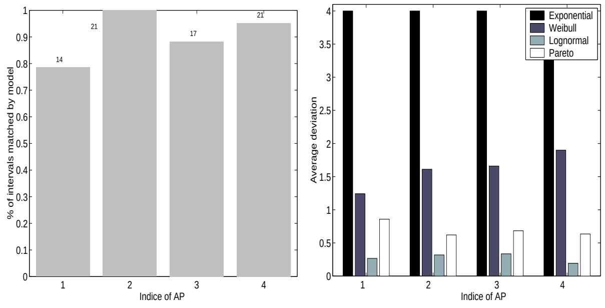
Next, we confirm that the flow size is best modeled by the Lognormal distribution. Figure 9.2 shows that such a distribution still yields the smallest average deviation which is less than 0.5. The result is consistent with the dataset Dartmouth02.

Unfortunately, the Stanford00 and IBM02 traces do not record the tcpdump data, we are thus unable to validate the models with those two scenarios in this section.

4.5 Summary

Each static flow can be characterized with three factors: its arrival time, its duration, and the idle periods during its lifetime. We further progressively model these factors at an AP and across several APs. In fact, the flow arrival process at each AP is nonstationary in coarse time granularity

[3] We use different time periods like 2 seconds and 10 seconds and obtain similar results



8.1: Accuracy of the Weibull regression model

8.2: Quality of fit for flow size

Figure 8: Modeling TCP connection arrivals and flow size by using the dataset UCSD01

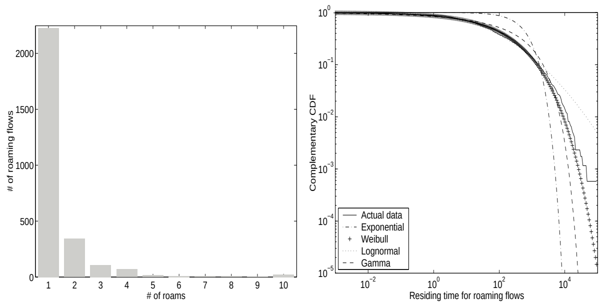
(say, several hours or days). However, it exhibits a dominant 24-hour periodic mode and a diurnal pattern within a day. At fine time scales (say, an hour), the heavy-tailed, Weibull distribution can accurately characterize flow arrivals. We then take a two-tier modeling approach to capture the features of both time scales. The result is a relatively simple, Weibull regression model that accurately approximates the flow arrival process in *all* time scales. For different APs, we further discover that, the parameters of the Weibull regression model are location dependent, and vary from one AP to another. However, APs in the same subnet observe spatial similarity in the sense that, the flow interarrival times across APs are highly likely to follow identical statistical distribution. As for the flow duration, we characterize it via the data size each flow transfers, and find that it follows the Lognormal distribution. Such a distribution holds for all APs. Regarding the idle period within a flow's lifetime, 97.3% flows remain non-idle all the time. For the remaining 2.7% flows, 66% idles only once.

The modeling results can be mostly explained by user behavior and application demands. The strong 24-hour periodic mode and diurnal pattern within a day attribute to human daily activity. The Weibull-alike arrival models the coexistence of short and long interarrival times, typical in popular applications like HTTP. Spatial similarity within a subnet captures the fact that local users tend to belong to the same interest group (e.g., a class) and share similar access patterns.

5. ROAMING FLOWS

In this section, we turn our attention into characterizing a unique feature of wireless networks, the *roaming flows*. Roaming flows transit more than one AP during their duration.

Our analysis shows that only 0.13% of all flows are actually roaming flows. Although the percentage is small, we believe that modeling roaming flows is important because such flows will incur resource dynamics in the spatial domain. How to manage related issues, such as assured service for handoff flows, admission control for flows, and fast handover, has been active research topics for years. Realistic modeling of roaming flows will benefit all such research.



9.1: Histogram for # of handoffs

9.2: Log-log CCDF for residing time

Figure 9: Modeling roaming flows

Host-level Roams	Have roaming flow	Have flow ending before roaming	Have flow starting after roaming
42077	5.8%	0.9%	1.9%

Table 4: Flow generation during host-level roams

Moreover, as applications like VoIP become more popular in wireless networks, the effect of roaming flows will become more visible.

5.1 Modeling Roaming Flows

Each roaming flow can be modeled by two factors: the number of handoffs and the residing time at each AP between consecutive handoff events. It turns out that the number of handoffs for a roaming flow fits the Geometric distribution very well; It passes the χ^2 test at the significance level $\alpha = 0.01$. The parameter p in the Geometric distribution is about 0.80 using the MLE technique. Figure 9.1 further shows the empirical histogram of the number of handoffs for roaming flows. It is interesting to observe that 97.7% roaming flows have less than 5 handoffs.

The good match of the Geometric distribution indicates that the handoff process for a roaming flow is memory-less. A roaming flow moves to another neighboring AP with probability p, and such a decision is independent of the past roaming history. This seems to confirm that the Exponential model (used in continuous domain) assumed in early analysis is pretty accurate.

We further show that, the flow residing time at each transit AP is best approximated by the Weibull distribution. It passes the χ^2 test at the significant level $\sigma = 0.05$. Figure 8.2 also shows that the Weibull distribution yields the best match in the log-log CCDF plot. The model indicates that the residing time the flow spends at each transit AP is heavy-tailed. It captures the bursty nature of the residing time: the residing time can be quite short when a host continually moves, but may become very long once the user slows down or even becomes stationary for an extended period of time.

5.2 Understanding Flows from Roaming Hosts

In order to further understand roaming flows, we further examine the host-level mobility behavior, since roaming flows are ultimately decided by host roaming patterns.

Port number	Roaming flows	Static flows
80 (http)	9.2%	29.0%
110 (pop3)	4.5%	30.5%
139 (netbios)	2.1%	15.6%
2151 (Blitzmail)	47.4%	6.4%
445 (microsoft-ds)	2.0%	1.9%
902 (ideafarm-chat)	4.4%	4.6%
1065 (syscomlan)	2.2%	0.08%

Table 5: Percentage of protocols for roaming flows and for static flows (this table only lists those protocols with percentage larger than 2%)

From the Dartmouth02 dataset, we find that about 60% hosts have host-level roaming events at least once. There are 51913 host-level roaming events recorded in the trace. However, only 77% are considered real events. The remaining 23% are caused by design flaws in hardware rather than physical movements [9]. Such events involved extremely frequent, back-and-forth inter-AP switching. We do not consider them in our analysis.

Among the roaming-host events, we find from Table 4 that in total only 8.6% of them really generate flow traffic, among which 5.8% host-level roams do carry roaming flows without tearing them down. Since a few host-level roams carry multiple roaming flows, these 5.8% host-level roams induce 2.8k roaming flows in total. We also observe that 0.9% host-level roams terminate ongoing flows before handoff and 1.9% hosts start new flows within 5 seconds after handoff.

We can see that though host mobility is pervasive and 60% hosts do handoff at least once, only 0.13% (2.8k out of the total 2.17 millon TCP flows) flows are roaming. The majority users, say, 91.4% in Dartmouth02 case, simply do not initiate or carry traffic while they are physically moving. We further examine the percentage of application protocols and find that the rarity of roaming flows may be due to the nature of applications at this time. As shown in Table 5, 59.5% of static flows are HTTP and POP3. However, these two applications contribute merely 13.7% to the roaming flows. The most popular protocol in the roaming flows takes port number 2151, which is used by the Dartmouth-specific email application Blitzmail.

Host mobility model As part of our measurement, we also model the host-level mobility. We first define *active period* as the time interval between a host associated with the WLAN and disconnected from the WLAN. Then we use three metrics to characterize the host-level mobility: the arrival process of active periods, the total number of transit APs visited within a single active period, and the residing time at each transit AP.

We present the results in Table 6. It turns out that the interarrival times of active periods are best modeled by the Lognormal distribution, which indicates that some interarrival times of the active periods are extremely long. Within each active period, the total number of APs visited is better modeled by the Gamma distribution, and the host residing time at each AP can be approximated by the Lognormal distribution. Therefore, both indicate that the handoff process for roaming hosts, which mainly reflects human behavior, is correlated, rather than independent of past history as many early studies assumed [8, 26].

	Dartmouth02	IBM02
Interarrival times	0.31 (Lognormal)	0.52 (Lognormal)
Residing time	0.25 (Lognormal)	0.43 (Lognormal)
# of APs visited	0.5 (Weibull)	2.68 (Gamma)

Table 6: Quality of fit for roaming active period

5.3 Cross-validation

We use the Stanford00 and IBM02 datasets to partially verify the results on roaming flows and roaming hosts. The validation is partial because these two traces do not log flow-level information. Stanford00 only records how each packet is transmitted by a host, while the IBM02 dataset does not have tcpdump records and contains only snmp data, which records the mobile user location every 5 minutes. Therefore, the best we can do is to identify host-roaming events and roaming hosts based on coarse estimations. From the Stanford00 dataset, we identify a host-roaming event, when two packets are transmitted from the same host and the same port number but via different APs, and the time gap between these two packets is less than 5 seconds. In the IBM02 dataset, a host-roaming event is assumed if a wireless user appears in different APs within 5 minutes.

We first find that the percentages of host-roaming events and roaming hosts are similar to those for the Dartmouth02 dataset. There exist 17350 host-roaming events in the Stanford00 dataset while only 11% of them generate packets, comparable to the Dartmouth02 case in which 8.6% of host-roaming events generate traffic. In the IBM02 dataset, about 69% of mobile users roam at least once while in Dartmouth02 the percentage is 60%. Since the Stanford00 trace does not provide precise flow information, we do not use it to verify the previous model of roaming flows.

As for host mobility model, the IBM02 dataset corroborates that for Dartmouth02, as shown in Table 6. Both interarrival times and residing times follow the Lognormal distribution.

6. APPLICATIONS OF DERIVED MODELS

This section discusses how to apply the derived models. We present two examples of scheduling and TCP to show how the flow-level models can be applied. The results show a 99.4% performance discrepancy in scheduling and a 7 times difference in the TCP performance gain, when comparing using previous models and using the ones presented in this work. We then discuss how the models can be used in more general context.

6.1 Case 1: Proportional Fair Scheduling

The first example is to demonstrate how to apply our models to help evaluate scheduling designs in a realistic setting. A recent award-winning paper [24] studied the flow-level performance of proportional fair scheduling. The author analyzed the flow-level[4] performance in terms of mean flow transfer delay (MFTD), blocking probability for new flows, etc. However, the result requires two idealized assumptions: (1) new flows arrive as a Poisson process, and (2) flow size is either fixed or exponentially distributed. Neither assumption is realistic according to our trace analysis.

[4] [24] uses *user-level* instead of *flow-level*. However, they convey the same meaning in the context of [24].

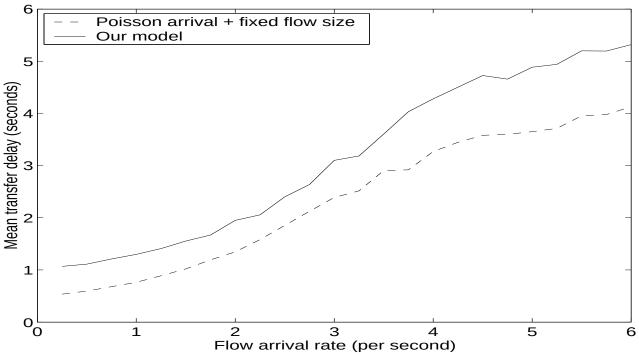


Figure 10: Comparing mean flow transfer delay (MFTD) between using the idealized flow models and the derived models

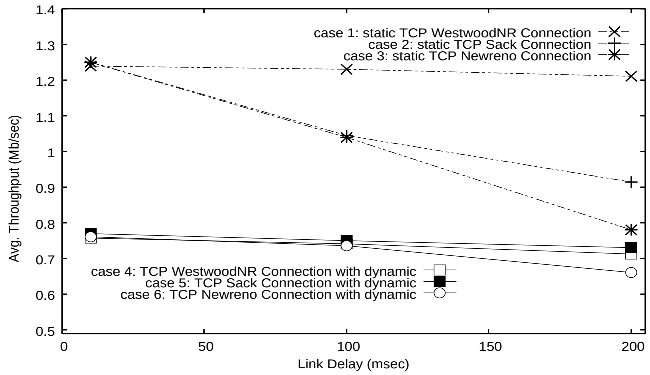


Figure 11: Comparing TCP performance between static and dynamic scenarios

To exemplify the performance discrepancy brought by using their idealized flow model and ours, we conducted a numerical experiment to plot MFTD versus the flow arrival rate (corresponding to Figure 2 in [24]). The simulated scenario is described as follows. New flows enter the system according to either the idealized flow model or our derived model. Each flow experiences an independent Rayleigh fading channel with the mean SNR being 0dB. The instantaneous SNR value is mapped to instantaneous data rate based on Table 1 of [24]. In each time slot, the system uses the Proportional Fair scheduling algorithm to schedule a flow. For either the idealized models or the derived models in this work, the mean flow arrival rate varies from 0.25 to 6 flows per second, while the mean flow size is fixed as 60 Kbytes.

We plot the results in Figure 10. The figure shows that as the mean flow arrival rate varies, the MFTD obtained by using our models is 24.1% to 99.4% larger than using the idealized models of [24]. In other words, the simulations using the idealized models produce consistently over-optimistic performance. In the worst case, such idealized models overestimate the MFTD by as high as 99.4%.

The significant increase in MFTD is not surprising. When compared with the Poisson arrivals and exponential or fixed size distribution, the derived Weibull regression and Lognormal distributions are both heavily tailed. Such heavy-tailed arrivals and flow size will lead to bursty flow arrivals and possibly very large flow size. Both factors will cause the mean waiting time to increase dramatically.

6.2 Case 2: TCP Performance in Wireless, Dynamic Scenario

In the second example, we use the derived models to measure TCP throughput in a more realistic, dynamic wireless setting. Most of the existing studies on TCP performance assume a static flow setting and unlimited source data for TCP transfer. We again show that simulations using such *static* configurations can produce results which are far from realistic situations and may be misleading. The projected 55% performance gain[5] in static settings may diminish to 7.8% when the flow configurations are dynamic.

We use *ns-2* [22] to measure TCP performance in both static and dynamic scenarios. The simulated topology is a simple wired-cum-wireless scenario consisting of a mobile node, a base-station node and a wired node. The mobile node sets up new TCP connections with the wired node via the base-station. There is only one TCP connection in the static case. In the dynamic case, the mobile node initiates new TCP connections according to the proposed Weibull regression model, and the size of the total data transmitted by each TCP connection follows the Lognormal model. We also vary the end-to-end RTT by letting the one-way propagation delay for the wired link vary from 10 ms to 200 ms. We measure three TCP protocols, i.e., TCP NewReno, TCP Sack and TCP Westwood [5], in both static and dynamic scenarios.

Figure 11 plots the throughput results for all three TCP protocols versus the propagation delay. In the static case, TCP Westwood performs best among all three, and shows a throughput gain of 55% over TCP NewReno when the one-way delay is 200ms. However, in the dynamic case using our derived models, all three TCP protocols perform much worse. The performance gain of TCP Westwood diminishes to only 7.8% in the static case. The explanation is that, in the dynamic setting, most of the flows are shorter-lived and experience slow-start phases. Thus, the link capacity is under-utilized and the aggregate throughput is much smaller when compared to the static setting.

6.3 Other Usage

In addition to performance evaluation of flow-level scheduling and TCP protocols, our results may also serve as an integrated component of a wireless system benchmark. Many other fields in computer systems, such as architecture, databases, and programming languages, have benefitted tremendously from a suite of benchmarks in design and evaluation of the system solutions. We feel that such a wireless system benchmark is long overdue. Conceptually, a wireless system benchmark has to cover network, application and mobile user aspects. Our derived flow-level models reflect the behavior of applications and users in the real world, and can be well integrated into the benchmark.

The derived models for roaming flows are also useful for admission control [25] and resource management [26], both of which require the knowledge of precise flow arrivals. In

[5] We were not able to reproduce the reported 550% throughput gain of [5], which did not use the wireless extension of ns-2, as remarked by the authors.

wireless networks, flow arrivals should contain not only static flows but also roaming flows. Since roaming flows may become popular in the future with the emerging of multimedia applications, our derived models for the roaming flows may benefit admission control and resource management.

7. DISCUSSION

Alternative modeling approach If we take the flow-level modeling approach, a related issue concerns with whether to model each application may yield better results. Essentially, flow-level modeling attempts to capture the impact of user demands and application behaviors. We do not take the approach of modeling each application directly. Our results show that our approach can yield an accurate match even only studying the aggregate effect of all applications and without examining each individual application. Another issue relevant to our flow arrival model concerns whether we should use multiple-tier instead of two-tier approach. While the model accuracy can be undoubtedly enhanced by taking the multiple-tier approach, more parameters are introduced and the resulting model may become complicated and unwieldy for applying. As we have shown earlier, the two-tier modeling approach strikes a balance between conciseness and accuracy. Therefore, we only consider two-tier approach in this work.

Measurement traces At the beginning of this work, we were quite excited to use four different traces that were collected from various distinctive settings of large campus, large corporate, major conference, and medium-sized department. Ideally, we would like to analyze all traces and cross-verify the obtained results. However, our validation effort in this work is only of partial success. The real roadblock is the incomplete trace record. For example, the IBM02 trace did not include any flow-level data. Therefore, we were unable to analyze and verify the flow-level characteristics in this network setting. We feel that in the future, we need a more systematic method to collect traces, which record all data at the flow and host levels. This can greatly benefit the future research on wireless network measurements.

8. CONCLUSIONS

In this paper, we have presented the first systematic study of flow-level analysis in large wireless data networks. In our study, we focus on two types of flows: static flows and roaming flows. We model each of them by using simple statistic distributions and explain the implications from the perspective of user behaviors and application demands. The results are obtained by analyzing a comprehensive dataset from a campus-wide wireless network, consisting of 2.17 million TCP flows on 1706 hosts. We further cross-validate our results using three different independently collected datasets, whenever feasible.

For static flows, we model the flow arrival process at each AP and the data size for these arrived flows. We find that though flow arrivals are highly nonstationary, they exhibit a significant 24-hour periodicity . We attribute this to the diurnal cycle of human activities. The flow arrival process can be well approximated by a simple, Weibull regression model with 5 parameters. The flow size is better modeled by the Lognormal distribution. In addition, we discover that flow arrivals in different APs within a single subnet exhibit strong similarity. This phenomenon seems to be caused by "clustered" user behaviors in a geographic proximity.

Next, we characterize a roaming flow by the number of handoffs and the residing time within each transit AP. The number of handoffs is found to follow the memory-less Geometric distribution. The residing time is well approximated by the Weibull distribution. Besides, we find that the occurrence of roaming flows is extremely rare despite the popularity of roaming hosts. This can be explained by the observation that few roaming hosts generate traffic when they are physically moving.

Our ultimate goal is to provide a flow-level model useful in aspects such as performance evaluation and benchmark design. We use two showcases to illustrate how they can be applied to evaluate flow-level scheduling algorithms and TCP performance in realistic traffic scenarios. In both showcases, we show that the performance has large discrepancy between using previously assumed models and using the models in this work. In the scheduling scenario, the performance gap can be as large as 99.4%. In the wireless TCP scenario, the performance gain may decrease from 55% to 7.8%. We see this work as a small step forward for providing realistic performance evaluation in wireless data network.

9. ACKNOWLEDGMENTS

We thank David Kotz and Kobby Essien for the Dartmouth02 dataset; Anand Balachandran and his coauthors for the UCSD01 dataset; Magdalena Balazinska and Paul Castro for the IBM02 dataset; Diana Tang and Mary Baker for the Stanford00 dataset. We particularly want to thank David Kotz for his efforts in delivering us the Dartmouth02 dataset, and also for his clarifications on a few questions regarding the dataset.

We gratefully acknowledge helpful discussions with Yingnian Wu and Songchun Zhu on statistical models. We also thank our group members of Wireless Networking Group (WiNG) and Internet Research Lab (IRL) at UCLA for their constructive comments. Special thanks to Petros Zerfos for his comments on many rounds of earlier drafts. We also gratefully acknowledge the anonymous reviewers for their invaluable criticisms.

10. REFERENCES

[1] R. 2616. Hypertext Transfer Protocol – http/1.1.

[2] A.Balachandran, G.M.Voelker, P.Bahl, and V.Rangan. Characterizing user behavior and network performance in a public wireless LAN. In *ACM SIGMETRICS*, 2002.

[3] A.Feldmann. Characteristics of TCP connections. In K.Park and W.Willinger, editors, *Self-similar Network Traffic and Performance Evaluation*, pages 367–399. John Wiley and Sons, 2000.

[4] A.Konrad, B.Y.Zhao, A.D.Joseph, and R.Ludwig. A markov-based channel model algorithm for wireless networks. *ACM Wireless Networks (ACM WINET Special Issue: Selected papers from MSWiM 2001)*, 9(3), 2003.

[5] C. Casetti, M. Gerla, S. Mascolo, M. Y. Sanadidi, and R. Wang. TCP westwood: Bandwidth estimation for enhanced transport over wireless links. In *ACM MOBICOM*, 2001.

[6] C.Barakat, P.Thiran, G.Iannaccone, C.Diot, and P.Owezarski. A flow-based model for Internet backbone traffic. In *ACM IMC*, 2002.

[7] Dataplot. http://www.itl.nist.gov/div898/handbook/index.htm.

[8] D.Hong and S.S.Rappaport. Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures. *IEEE Trans. on Veh. Technol.*, 3:77–92, 1986.

[9] D.Kotz and K.Essien. Analysis of a campus-wide wireless network. In *ACM MOBICOM*, 2002.

[10] D.Kotz and K.Essien. Analysis of a campus-wide wireless network. to appear in *ACM Mobile Networks and Applications (MONET)*, 2004.

[11] T.Henderson, D.Kotz and I.Abyzov. newblock The changing usage of a mature campus-wide wireless network. In *ACM MOBICOM*, 2004.

[12] D.P.Heyman, T.V.Lakshman, and A.L.Neidhardt. New method for analyzing feedback protocols with applications to engineering web traffic over the Internet. In *ACM SIGMETRICS*, 1997.

[13] D.Schwab and R.Bunt. Characterising the use of a campus wireless network. In *IEEE INFOCOM*, 2004.

[14] D.Tang and M.Baker. Analysis of a metropolitan-area wireless network. In *ACM MOBICOM*, 1999.

[15] D.Tang and M.Baker. Analysis of a local-area wireless network. In *ACM MOBICOM*, 2000.

[16] F.Chinchilla, M.R.Lindsey, and M.Papadopouli. Analysis of wireless information locality and association patterns in a campus. In *IEEE INFOCOM*, 2004.

[17] G.D.Veciana, T.J.Lee, and T.Konstantopoulos. Stability and performance analysis of networks supporting services with rate control - could the Internet be unstable. In *IEEE INFOCOM*, 1999.

[18] J.Cao, W.S.Cleveland, D.Lin, and D.X.Sun. On the nonstationarity of Internet traffic. In *ACM SIGMETRICS*, 2001.

[19] J.Devore and R.Peck. *Statistics: the Exploration and Analysis of Data*. West Publishing Company, 1986.

[20] L.Massoulie and J.W.Roberts. Bandwidth sharing and admission control for elastic traffic. *Telecommunication Systems*, 15:185–201, 2000.

[21] M.Balazinska and P.Castro. Characterizing mobility and network usage in a corporate wireless local-area network. In *ACM MOBISYS*, 2003.

[22] ns2 network simulator. http://www.isi.edu/nsnam/ns/.

[23] S.B.Fredj, T.Bonald, A.Proutiere, G.Regnie, and J.Roberts. Statistical bandwidth sharing: A study of congestion at flow level. In *ACM SIGCOMM*, 2001.

[24] S.Borst. User-level performance of channel-aware scheduling algorithms in wireless data networks. In *IEEE INFOCOM*, 2003.

[25] S.Choi and K.G.Shin. Comparison of connection admission-control schemes in the presence of hand-offs in cellular networks. In *ACM MOBICOM*, 1998.

[26] S.Lu and V.Bharghavan. Adaptive resource management algorithms for indoor mobile computing environments. In *ACM SIGCOMM*, 1996.

[27] S.Sarvotham, R.Riedi, and R.Baraniuk. Connection-level analysis and modeling of network traffic. In *ACM IMW 2001*, 2001.

[28] V.Paxson. Empirically derived analytic models of wide-area TCP connections. *IEEE/ACM Transactions on Networking*, 2(4):316–336, 1994.

[29] V.Paxson and S.Floyd. Wide-area traffic: The failure of poisson modeling. *IEEE/ACM Transactions on Networking*, 3(1):226–244, 1995.

11. APPENDIX

11.1 Summary of Distribution Models

Model	PDF $f(x)$	CCDF $P(X > x)$
Exponential	$\frac{1}{\mu}e^{-\frac{x}{\mu}}(x > 0)$	$e^{-\frac{x}{\mu}}$
Weibull	$\mu k x^{k-1}e^{-\mu x^k}(x > 0)$	$e^{-\mu x^k}$
Lognormal	$\frac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{(logx-\mu)^2}{2\sigma^2}}(x > 0)$	$\phi(\frac{logx-\mu}{\sigma})$
Gamma	$\frac{\lambda(\lambda x)^{k-1}e^{-\lambda x}}{\Gamma(k)}(x > 0)$	$1 - \frac{\int_0^{\lambda x} u^{k-1}e^{-u}du}{\Gamma(k)}$
Pareto	$ak^a x^{-a-1}(x \geq k)$	$(\frac{k}{x})^a$
Extreme -value	$\frac{1}{\beta}e^{-\frac{x-\alpha}{\beta}}e^{-e^{-\frac{x-\alpha}{\beta}}}$	$1 - e^{-e^{-\frac{x-\alpha}{\beta}}}$
Geometric	$(1-p)^{x-1}p(0 < p < 1)$	$(1-p)^x$

Table 7: Seven analytical distributions ($\phi(.)$ is the standard normal distribution. $\Gamma(.)$ is the gamma function)

11.2 Estimating Parameters in the Weibull Regression Model

There are five parameters involved in the Weibull regression model, i.e., k, t_h, t_l, μ_0 and β. k is estimated by using a two-parameter Weibull model over the entire set of flow arrivals. t_h, t_l is estimated by using a simple heuristic threshold-based method. More specifically, if the arrival rate keeps increasing/decreasing in a continuous three-hour period and the increased/decreased amount is larger than half of the mean arrival rate, then the starting time of the three-hour period is assigned to t_h (t_l). To estimate μ_0 and β, we use Maximum Likilihood Estimation (MLE). In the following, we describe the estimation method in more details.

Let $y = logx$, we rewrite (1) as

$$f(y|t) = k\mu(t)e^{ky-\mu(t)e^{ky}} \tag{3}$$

Now suppose there are n samples of interarrival time observed from real traces. The i-th sample is denoted by x_i and there is $y_i = logx_i$. The time-of-day for x_i is denoted by t_i, thus the likelihood function for all the n samples is

$$L(\mu_0, \beta) = \prod_{t_i \in T_h} k\mu_0 e^{ky_i - \mu_0 e^{ky_i}} \prod_{t_i \in T_l} k\mu(t_i)e^{ky_i - \mu(t_i)e^{ky_i}}$$

which is equivalent to

$$\begin{aligned}\log L(\mu_0, \beta) &= \sum_{\forall i}(\log k + ky_i) + \sum_{t_i \in T_h}(\log \mu_0 - \mu_0 e^{ky_i}) \\ &+ \sum_{t_i \in T_l}[\log \mu(t_i) - \mu(t_i)e^{ky_i}]\end{aligned}$$

After substituting (2) into the above equation,we get the first-order derivatives of $\log L(\mu_0, \beta)$

$$
\begin{aligned}
\frac{\partial \log L(\mu_0, \beta)}{\partial \mu_0} &= \sum_{t_i \in T_l} \frac{1}{\mu_0 - \beta(t_i - t_l)(t_h + 24 - t_i)} \\
&+ \frac{n}{\mu_0} - \sum_{\forall i} e^{k y_i} \\
\frac{\partial \log L(\mu_0, \beta)}{\partial \beta} &= -\sum_{t_i \in T_l} \frac{(t_i - t_l)(t_i - t_h - 24)}{\mu_0 - \beta(t_i - t_l)(t_h + 24 - t_i)} \\
&+ \sum_{t_i \in T_l} (t_i - t_l)(t_h + 24 - t_i) e^{k y_i}
\end{aligned}
$$

The maximum likelihood equations $\frac{\partial \log L(\mu_0,\beta)}{\partial \mu_0} = 0$ and $\frac{\partial \log L(\mu_0,\beta)}{\partial \beta} = 0$ enable us to rewrite the above two equations as

$$
\begin{aligned}
\sum_{t_i \in T_l} \frac{1}{1 - \varepsilon(t_i - t_l)(t_h + 24 - t_i)} &= \mu_0 \sum_{\forall i} e^{k y_i} - n \qquad (4) \\
\sum_{t_i \in T_l} \frac{(t_i - t_l)(t_i - t_h - 24)}{1 - \varepsilon(t_i - t_l)(t_h + 24 - t_i)} &= \mu_0 \sum_{t_i \in T_l} (t_i - t_l)(t_h + \\
& 24 - t_i) e^{k y_i} \qquad (5)
\end{aligned}
$$

where $\varepsilon = \frac{\beta}{\mu_0}$.

To estimate the unknown parameters ε and μ_0, we first merge (4) and (5) to remove μ_0, we then readily apply Newton-Raphson method (or some other iterative methods) to the resulting equation to estimate $\hat{\varepsilon}$. Once $\hat{\varepsilon}$ is known, the MLE $\hat{\beta}, \hat{\mu}_0$ are also solved. Note that the convergence of Newton-Raphson method is sensitive to the initial values chosen. Based on our experience, the initial value for $\hat{\varepsilon}$ should be chosen as $\max_i \frac{1}{(t_i - t_l)(t_h + 24 - t_i)} + \epsilon$, in which ϵ is a small positive value, e.g., 10^{-6}.