Spatio-Temporal Modeling of Traffic Workload in a Campus WLAN

Félix Hernández-Campos*
Department of Computer Science
University of North Carolina
Chapel Hill, United States
fhernand@cs.unc.edu

Merkouris Karaliopoulos
Department of Computer Science
University of North Carolina
Chapel Hill, United States
mkaralio@cs.unc.edu

Maria Papadopouli†
Department of Computer Science
University of Crete
Heraklion, Greece
maria@csd.uoc.gr

Haipeng Shen
Department of Statistics and Oper. Research
University of North Carolina
Chapel Hill, United States
haipeng@email.unc.edu

ABSTRACT

Campus wireless LANs (WLANs) are complex systems with hundreds of access points (APs) and thousands of users. Their performance analysis calls for realistic models of their elements, which can be input to simulation and testbed experiments but also taken into account for theoretical work. However, only few modeling results in this area are derived from real measurement data, and rarely do they provide a complete and consistent view of entire WLANs. In this work, we address this gap relying on extensive traces collected from the large wireless infrastructure of the University of North Carolina. We present a first system-wide, multi-level modeling approach for characterizing the traffic demand in a campus WLAN. Our approach focuses on two structures of wireless user activity, namely the wireless session and the network flow. We propose statistical distributions for their attributes, aiming at a parsimonious characterization that can be the most flexible foundation for simulation studies. We simulate our models and show that the synthesized traffic is in good agreement with the original trace data. Finally, we investigate to what extent these models can be valid at finer spatial aggregation levels of traffic load, *e.g.*, for modeling traffic demand in hotspot APs.

*Dr. Hernández-Campos is currently with Google Inc.

†Prof. Papadopouli is also affiliated with the Institute of Computer Science, Foundation for Research and Technology -Hellas, Greece, and the Department of Computer Science, University of North Carolina, Chapel Hill, NC, United States.

TAPAS'06, First International Workshop on Technology and Policy for Accessing Spectrum, August 5, 2006, Boston, MA, United States

Categories and Subject Descriptors

I.6 [**Simulation and Modeling**]: Model Validation and Analysis; I.6.6 [**Model Development**]: Modeling methodologies

General Terms

Measurement, Experimentation

1. INTRODUCTION

Wireless local area networks (WLANs) are increasingly being deployed to address the growing demand for wireless access. For the support of real-time multimedia services, capacity planning, link adaptation, load balancing are amongst the mechanisms that have to be deployed to provide a better than-best-effort service. For their performance analysis, models of the network and user activity are critical.

One of the most intriguing aspects of the traffic demand modeling task in WLANs is its multi-level spatio-temporal nature, namely the different spatial scales (*e.g.*, infrastructure-wide, AP-level or client-level) and time granularities, such as packet-level, flow-level and session-level, inherent in the task. Key structures of this demand are the WLAN client associations and the traffic flows. We study client association dynamics using sessions, which group client associations into episodes of continuous activity. The *s*ession-level captures the interaction between clients and the network infrastructure and is fundamental for the study of mechanisms that maintain state in APs. The flow-level is an important structure above the packet-level for network traffic analysis and closed-loop traffic generation. How do clients arrive at an AP or in the campus-wide infrastructure? How are flows generated at APs? What are their temporal dynamics? Sessions and flows are interrelated: the load of an AP is given by the set of network flows that traverse this AP, generated by the clients associated to it. This paper models these structures in both spatial and temporal dimensions and investigates their dependencies and interrelation.

Whereas there is rich literature on traffic characterization in wired networks (*e.g.*, [25, 4, 7, 6, 18]), there is significantly less work of the same detail for WLANs. Hierarchical approaches to modeling the wireless demand and its spatial and temporal phenomena have received little attention from our community. In fact, the only relevant study we are aware of is the flow-level modeling study by Meng *et al.* [17]

The first contribution of this paper is methodological in that it models the demand in large wireless networks taking a system-wide, multi-level parametric approach. Our approach distinguishes two important dimensions in wireless network modeling, namely the user demand (user-initiated activity through flows and sessions) and the topology (network, infrastructure, and radio propagation dependencies). This enables us to "superimpose" models for the demand on a given topology, and focus on the right level of detail for the performance analysis or simulation study (*e.g.*, AP-level, system-wide, client-level). This methodology "masks" network-related dependencies that may not be relevant to a range of systems and makes the wireless networks amenable to statistical analysis and modeling. To the best of our knowledge, this is the first system-wide multi-level modeling study of traffic demand in WLANs.

Besides the methodological aspects of our work, our main contribution consists of coherent parametric statistical models of the workload of the entire WLAN. Our parsimonious description of the workload seems very appropriate for simulation and testbed experimentation studies, while it allows better insight to the problem than empirical models. The network load can be simulated at both the client association and flow levels by using models of the compound process of sessions and flows. As we show, sessions have a well-behaved arrival process, which can be accurately described using a time-varying Poisson process. In addition, an AP preference distribution can be used as a first rough approximation for distributing sessions throughout the wireless infrastructure in a manner that is representative of real workloads. The session arrival process provides the seeds for a cluster process, in which the arrivals of sessions imply the arrivals of correlated sets of flows. Simulations can first produce a time series for the session arrival process, and then sample the distributions of the number of flows and their inter-arrivals to generate the within-session flow arrival time series. The simulation assigns a flow size to each flow based on the proposed distribution. Packet-level details are left to the underlying protocols and are beyond the scope of our modeling work.

Our contributions are summarized as follows:

- A methodology for the statistical modeling of wireless network traffic demand, relying on robust statistical methods to study large-scale phenomena.
- System-wide and AP-level models of traffic demand. They are more intuitive and parsimonious than the ones in [17], and capture the network-independent characteristics of the traffic workload.
- Validation of our modeling results showing their agreement with the measurement data.

The next section briefly reviews the wireless infrastructure at the University of North Carolina (UNC) and data acquisition process. Section 3 describes our overall modeling methodology. Our modeling results are presented and evaluated in the next two sections. Section 4 considers the spatio-temporal characteristics of the entire system, whereas Section 5 compares our model-driven synthetic traffic with the original traces. We test the applicability of the proposed system-wide traffic models for modeling traffic demand in hotspot APs in Section 6. Section 7 positions our study with respect to related work in literature, and Section 8 summarizes our main results and future work plan.

2. WIRELESS INFRASTRUCTURE AND DATA ACQUISITION

Our data come from the large wireless network infrastructure of the UNC campus. By the time the measurements were made, about 500 APs provided wireless access to 26,000 students, 3,000 faculty members and 9,000 staff members all over the 729-acre campus and a couple of off-campus administrative offices. The covered building types vary widely: from academic buildings and libraries to student dormitories and sport halls.

The majority of APs belong to the Cisco 1200 Aironet series; the network also features a significant number of 350 series APs and fewer 340 series APs. Two are the main trends with respect to the infrastructure evolution with time: it is constantly growing, with APs exceeding 750 by June 2006 and, in parallel, older 340/350 series APs tend to be replaced by 1230/1240 AG series APs [1].

Two types of measurement data have been used in this study. SNMP data are collected from each AP every five minutes. We use a custom data collection system, being careful to avoid the pitfalls described in [10]. The system relies on a non-blocking SNMP library for polling APs in an independent manner and eliminating any extra delays due to the slow processing of SNMP polls by some of the slower APs. SNMP polling has been carried out continuously from September 29th, 2004 until June 26th, 2005. The monitoring system did not suffer any problems during this period.

However, our analysis concentrates on an 8-day period, from 12:06 PM on Wednesday April 13rd, 2005 till 22:18 PM on Wednesday April 20th, 2005, over which we also collected wireless traffic flow data. Our 178.2 hour long data set consists of a total of 175 GB of packet header traces captured on the link between UNC and the rest of the Internet. The packet headers were acquired using a high-precision monitoring card (Endace DAG 4.3 GE) attached to the receiving end of a fiber split. The card was installed in a high-end FreeBSD server. Neither the server nor the card's driver reported any failures or packet drops during the monitoring process.

The SNMP data are cross-compared with the packet header trace data and the timestamps in the two datasets are used to extract the time bounds of the client sessions and identify the traffic flows that were initiated in each one of them. The focus in this study is on TCP connections, which constitute the vast majority of the captured traffic.

Our initial intention was to also examine datasets from the Dartmouth University campus, in continuation of the higher-level comparative study of the two networks in [13]. Unfortunately, the available data from the Dartmouth campus do not allow a direct comparison; packet header traces are collected by a subset of its wireless infrastructure (31 APs), while the collected SNMP data do not include all information required for our two-level modeling approach.

3. MODELING METHODOLOGY

Our modeling approach draws on two fundamental concepts, the *wireless session* and *network flow*.

A wireless session can be viewed as an episode in the interaction of a client and the wireless infrastructure: a wireless client arrives at the network, associates to one or more APs for some period of time, and then leaves the infrastructure. As we will demonstrate, sessions are statistically well-behaved, and, most significantly, robust to network dependencies. There is consensus in the network community that traffic modeling should not address elements that are dominated by too specific network-side characteristics or conditions. Otherwise, simulations and experiments using the respective models can *never* study changes in those conditions or new network mechanisms that shape those conditions. For example, in the context of WLANs, modeling the precise sequence of associations and disassociations inside sessions is too network-specific, since small changes in the network layout, physical environment, or network/client equipment can dramatically change association/disassociation dynamics. A new proposed algorithm for AP selection may also change association dynamics. Therefore, the simulation model should not impose *a priori* a certain sequence of associations and disassociations. This requirement is satisfied when sessions are the subject of modeling. The simulated session may end up having completely different association dynamics, but the corresponding workload (i.e., generated traffic during a time period) is preserved.

In our approach, sessions represent the high level unit of wireless network traffic load, including all the packets sent and received by the APs due to the client's communication with one or more Internet hosts. On the other hand, network flows provide a finer level of modeling the packet-level workload. Working with flows, such as TCP connections and UDP conversations, is in line with the approach taken in [18, 17, 22] and the principles of network-independent modeling from [23]. Network flows are well-separated collections of packets between a pair of Internet hosts, *i.e.*, packets that share the same transport-layer "5-tuple". In our model, a session groups the set of flows started by a client. Therefore, simulating the traffic workload consists of simulating sessions and the flows started inside them, leaving packet-level and association dynamics to underlying mechanisms that are independent of our model.

We have chosen to rely on parametric models for the traffic demand variables. When compared with empirical models, they provide better insight to the properties and the dynamics of the modeled quantities. In parallel, they are more adequate in summarizing datasets and make their comparison straightforward. Therefore, we propose statistical distributions for both session- and flow-level traffic variables. Particularly relevant in this context is the biPareto distribution, proposed in [18] to model the number of TCP connections per HTTP user session and the average inter-connection time within a session. In deriving distributions that best fit our data, we repeatedly make use of formal and visual statistical analysis methods and tools, such as the quantile plots with simulation envelopes. The interested reader may find more details regarding these plots in [11] and about the biPareto distribution in [18] and the Appendix.

4. SYSTEM-WIDE MODELING OF TRAFFIC DEMAND

4.1 Session Arrivals

The starting point of our model is the process of session arrivals. Figure 1 shows the point process of session arrivals for the 8-day trace. Each dot in the scatterplot corresponds to the arrival of a session, and each arrival is placed according to its temporal (arrival time in x-axis) and its spatial (AP of first association in the y-axis) coordinate. Although session arrivals vary widely, some expected patterns are apparent. Firstly, there is a clear diurnal periodicity, which is related to the substantial decrease of the network activity during the nights. Secondly, the activity of network clients decreases during the weekend (days 3 and 4 in the plot). These temporal patterns appear to be common throughout the AP population, although some APs are more likely to be used at night than others.

Figure 2 provides an even clearer picture of these temporal variations. It plots the time-series of session arrivals for the entire network using 1-hour bins. The time-series plot shows sharp increase in the number of session arrivals in the morning, reaching a peak between 1,000 and 1,110 sessions per hour during weekdays and 350 session arrivals per hour during the weekend. This pattern generally holds throughout the ten months covered by our SNMP dataset, except for specific time periods, such as the Christmas break, during which the activity decreases considerably and the diurnal/weekly variation is milder.

Figure 1 also lets some insight to the spatial dimension of the session arrival process, *i.e.*, the way sessions arrivals are distributed amongst the network APs. Although the AP ranking along the y-axis is random, it clearly hints at wide spatial variability of the workload. To illustrate this clearly, Figure 3 plots the probability that a session is initiated at a given AP, hereafter called *AP-preference* distribution. APs in this plot are numbered in order of their popularity as session-starting points, lower indices indicating more popular APs. The plot suggests that a few APs receive a substantial fraction of all sessions, with most APs being the starting points for only a few wireless sessions.

One remarkable aspect of Figures 2 and 3 is the smooth-

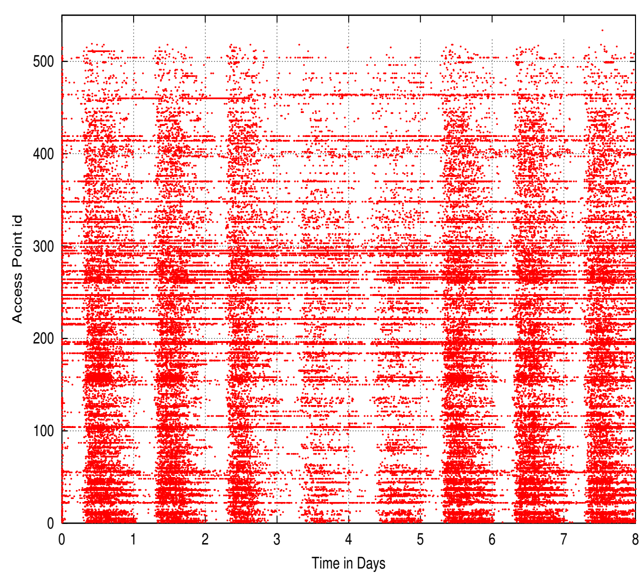

Figure 1: Arrivals of sessions from wireless clients over time and across the campus APs.

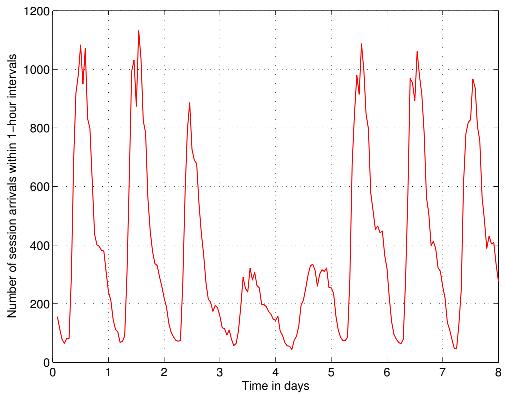

Figure 2: Time-series of session arrivals in the entire campus WLAN (1-hour bins).

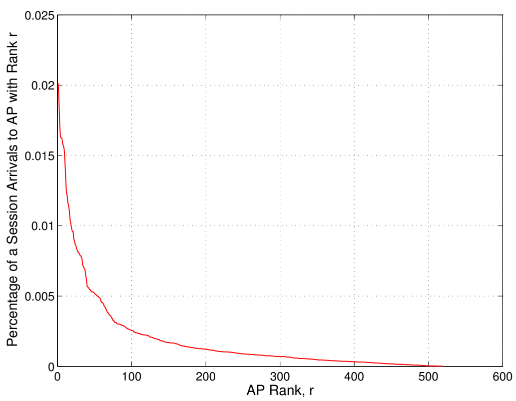

Figure 3: AP-preference distribution: APs are sorted by decreasing popularity.

ness of the curves, suggesting phenomena that are amenable to modeling. In fact, our analysis reveals that session arrivals follow a time-varying Poisson process, and that the AP-preference distribution is accurately described by a lognormal distribution.

4.1.1 Session arrival process

We model the session arrival process as a time-varying Poisson process and test the validity of our modeling assumption with the statistical test described in the Appendix. For the model to be valid, the variables R_{ij}s, which are defined in (1) as functions of the ordered session arrival times, must be exponentially distributed with a mean equal to unity and uncorrelated. The top part of Figure 4 shows an exponential quantile plot of the R_{ij}s during one randomly chosen hour.

We set the block length $L = 0.1$ hours in calculating the R_{ij}s. The red quantile plot follows closely the green diagonal line and remains well within the blue simulation envelope. This suggests that the exponential fit is clearly appropriate. The maximum likelihood estimate of the exponential parameter is 0.9372, which is very close to unity, and agrees with the claim that the R_{ij}s are standard exponential. The bottom plot of the figure plots the autocorrelations of the R_{ij}s up to 20 lags. The sample autocorrelations are always within

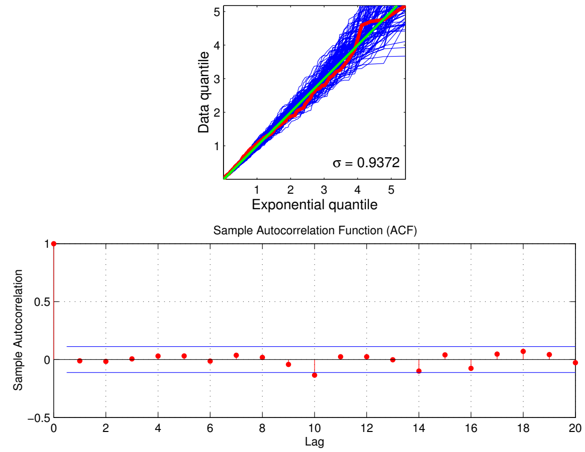

Figure 4: The R_{ij}s are independent and exponentially distributed. Only one hourly block is shown here, but the results are consistent across the entire dataset.

the confidence intervals, so the R_{ij}s do not exhibit any significant correlations. We got similar results when repeating the same analysis for other one-hour intervals of the 8-day dataset.

4.1.2 AP-preference distribution

Our analysis shows that a lognormal distribution with parameters $\mu = 4.0855$ and $\sigma = 1.4408$ is a good model for the AP preference distribution. As we can see in Figure 5, the original data, shown in red, lie within the natural variability of the lognormal model, since they remain within the blue simulation envelope. The only departure from lognormality is for the smallest values, *i.e.*, for APs that more rarely serve as session-starting APs, hence featuring very small number of samples. Overall, the lognormal distribution is an excellent description of the data. We have also considered other models but they are clearly outperformed by the lognormal fit. For example, Zipf's law, a classic way of describing popularity, is very far from the AP-preference distribution in our data.

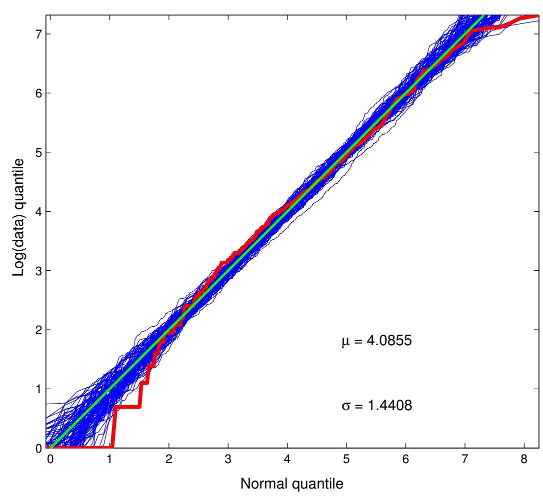

Figure 5: Lognormal model of AP-preference distribution.

4.2 Flow Arrival Process

At the next modeling level, the arrival of a session triggers the arrival of a group of flows, initiated between the client and one or more Internet hosts. It is therefore natural to describe flow arrivals as a cluster process [18] rather than a point process in which flows arrivals are described in isolation. Since session arrival counts are (time-varying) Poisson distributed, flow arrivals form a cluster Poisson process. The flow-level traffic variables that need to be modeled with this approach are the number of flows associated to each session-cluster, and the inter-arrivals of flows within sessions.

4.2.1 *Number of flows within session*

Our analysis showed that the biPareto distribution yields the best fit for the number of flows per session. Figure 6 plots the complementary cumulative distribution function of the fitted distribution against the empirical data in a logarithmic scale.

The red circles are an equidistant set of samples from a biPareto distribution with parameters $\alpha = 0.06, \beta = 1.72, c = 284.79$ and $k = 1$. The empirical distribution of the number of flows (in blue) matches well our model for probabilities between 0 and 0.995. The fit is worse at the tail due to sampling artifacts. In any event, it is clear that the biPareto model fits the empirical distribution very well.

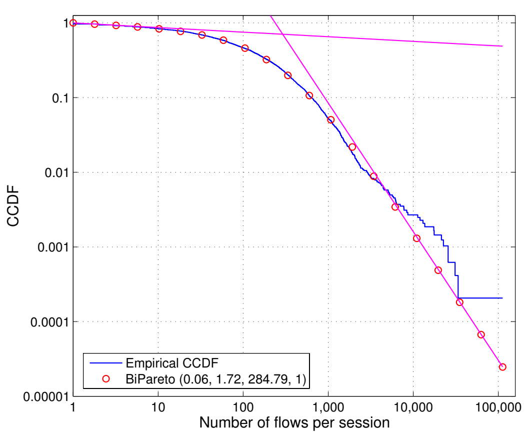

Figure 6: Number of flows per session.

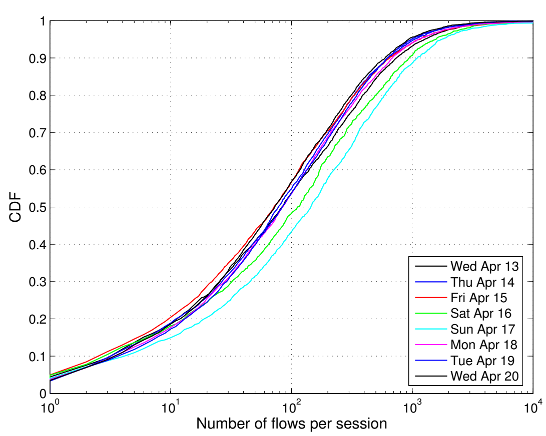

Figure 7: Stationarity of the distribution of the number of flows per session (body).

We have also studied how the distribution of the in-session number of flows varies per day. Figure 7 plots the distribution of the variable for each one of the 8 days in the dataset (see [11] for a plot of the ccdf). The eight distributions are very similar, with the vast majority of the sessions having between 1 and 1000 flows. The distributions for the weekends are slightly heavier. The number of flows per session goes as far as 10,000 for 0.1% of the sessions. This striking consistency of the eight curves strongly indicates that it is feasible to use parametric models for the traffic variables.

4.2.2 *Flow interarrivals within session*

The second component of our cluster model is the distribution of the flow inter-arrivals within sessions. We show that a lognormal model provides the best fit, although the distribution is rather complex. The lognormal quantile plot for the empirical data is shown in Figure 8; the parameters are estimated to be $\mu = -1.3674$ and $\sigma = 2.785$ using maximum likelihood. The red quantile plot follows the green diagonal line closely for all of the quantiles. The simulation envelope is very narrow in this case, and shows that some deviations from the lognormal model in the upper part are significant. While more complex models, *e.g.*, an ON/OFF model, may provide a better approximation, our lognormal fit certainly provides a reasonable description of the data using only two parameters.

We have also studied the stationarity of the flow inter-arrivals within sessions. Both their cdf (Figure 9) and ccdf (see [11]) plots suggest that the flow inter-arrivals during each day are very consistent with each other.

4.3 Flow Sizes and Packet-Level Load

To enable generation of the packet-level load in a manner suitable for closed-loop simulation and testbed experimentation, it is necessary to describe not only the flow arrival process but also the flow sizes in terms of number of bytes they transfer. Our statistical analysis reveals that flow sizes can be accurately described using a biPareto distribution

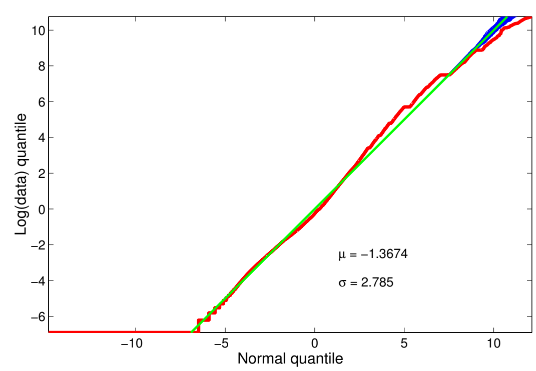

Figure 8: Flow inter-arrivals within a session: lognormal quantile plot of the data with a simulation envelope.

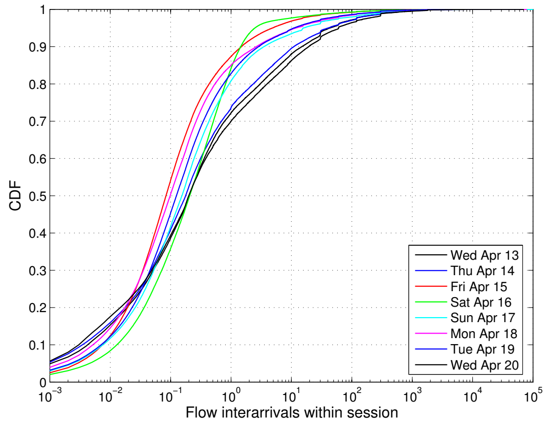

Figure 9: Stationarity of the distribution of flow inter-arrivals within sessions (body).

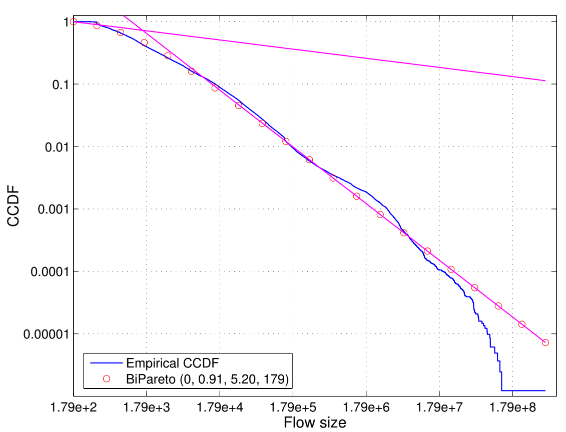

Figure 10: BiPareto model of flow sizes.

with parameters $\alpha = 0.00, \beta = 0.91, c = 5.20$ and $k = 179$. Figure 10 plots the biPareto fit (red circles) to the empirical data (blue curve). The fit is excellent for most of the distribution with biPareto clearly capturing the transition in the slope between the body and the heavy tail of the empirical distribution. The approximation appears heavier than the empirical data at the end of the tail, which could motivate further refinements of the fit.

We have also examined the stationarity of the flow size distributions over different days (the respective plots are provided in [11]). We found consistent tails for the eight days suggesting that weekly periodicities are not critical for modeling the flow sizes.

Table 1 summarizes our proposed statistical models for the system-wide traffic workload.

5. MODEL VALIDATION

5.1 Methodology

We evaluate the efficiency of our proposed system-wide models via simulation. The synthetic traffic generated according to the models described in Section 4 is compared against the original trace. Furthermore, we synthesize traffic via simulation of two other modeling alternatives. The first one is the compound model described in [18], which also discriminates between sessions and flows but differentiates in the way the within-session flow interarrivals are modeled. We refer to this model as the *compound* model in the subsequent discussion and plots. The second method is the flat flow-level modeling approach, where there is no session concept. The flow arrival process is assumed to be a renewal process; we estimate the empirical distribution of flow-interarrivals directly from the trace and use it to generate the time series of flow arrivals in the synthetic traffic generator. We simulate this model only as a comparison reference, to better illustrate the advantages of the two-level approach.

Given the heavy-tailed session duration, we impose simulation times in the order of days. In particular, we let the simulator synthesize traffic over a three-day interval (simulation time) and process the measured traffic variables obtained in the third day. To simulate the time-varying Poisson process for the per-hour session arrivals, which is required for our two-level model and the compound model, we use the thinning process described in [14].

In order to validate the model, we consider traffic variables *not* explicitly addressed by our models. Such variables are the aggregate flow arrival count process and the aggregate flow interarrival time-series. For the former, we plot the nunber of aggregate flow arrivals with time and their Coefficient of Variation (CoV) when estimated over different time scales. For the aggregate flow interarrivals, we examine the first-order (quantile plot) and second-order (autocorrelation function) statistics.

5.2 Aggregate flow interarrivals

We plot the quantiles of the simulated data from our model against the original trace data in Figure 11. The match is excellent and only for values exceeding the 99.9th percentile of simulated data do we see some deviation between the two datasets. The compound model of Nuzman *et al.* [18] performs worse (see [11]). Note that we have found that the flow interarrivals within a session follow a lognormal distribution; the compound model with the transformed Weibull variables cannot give an equally good fit for these interarrivals and this is reflected in the aggregate flow interarrival data.

Figure 12 plots the autocorrelation function of the synthetic aggregate flow interarrivals as estimated from our simulated model against the original trace. Though less precise than with first-order statistics, the simulated curve implies that the model can capture the second-order dynamics in the trace.

5.3 Aggregate flow arrivals

Figure 13 depicts the number of aggregate flow arrivals within intervals of one hour. The two-level model tracks closely the original trace in this respect, and certainly better than the other two approaches, although it overestimates the arrivals during the busy hours. The compound model yields less satisfactory matching, although it can respond to the non-stationarity of flow arrivals thanks to its provision for time-varying Poisson session arrivals. On the contrary, the flat model cannot respond to the time variations of flow arrivals, since the empirical distribution is estimated over the full trace and averages the hourly fluctuations of the traffic demand.

Table 1: Summary of models for system-wide traffic demand variables.

Modeled variable	Model	Probability Density Function (PDF)	Parameters
Session arrival	Time-varying Poisson with rate $\lambda(t)$	N: # of sessions between t_1 and t_2 $\lambda = \int_{t_1}^{t_2} \lambda(t)dt$, $Pr(N=n) = \frac{e^{-\lambda}\lambda^n}{n!}$, $n = 0, 1, \ldots$	Hourly rate: 44 (min), 1132 (max), 294 (median)
AP of first association/session	Lognormal	$p(x) = \frac{1}{\sqrt{2\pi}x\sigma} \exp\left[-\frac{(\ln x-\mu)2}{2\sigma 2}\right]$	$\mu = 4.0855, \sigma = 1.4408$
Flow interarrival/session	Lognormal	Same as above	$\mu = -1.3674$, $\sigma = 2.785$
Flow number/session	BiPareto	$p(x) = k^{\beta}(1+c)^{\beta-\alpha}x^{-(\alpha+1)}(x+kc)^{\alpha-\beta-1}$ $(\beta x + \alpha kc)$, $x \geq k$	$\alpha = 0.06, \beta = 1.72,$ $c = 284.79, k = 1$
Flow size	BiPareto	Same as above	$\alpha = 0.00, \beta = 0.91,$ $c = 5.20, k = 179$

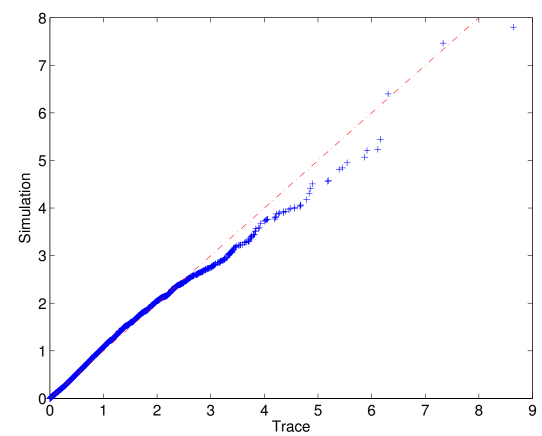

Figure 11: Quantile-quantile plot of the aggregate flow interarrivals: simulated two-level model vs. trace.

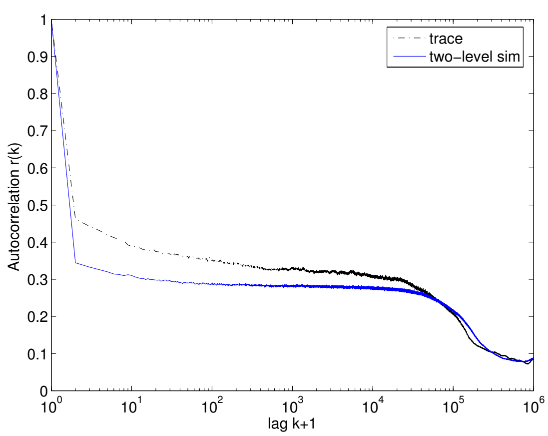

Figure 12: Autocorrelation of aggregate flow interarrivals: simulated two-level model vs. trace.

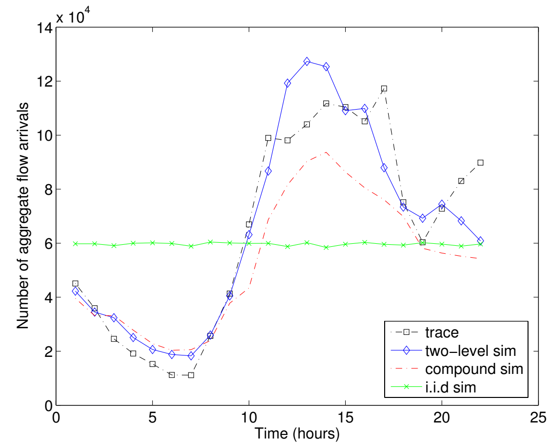

Figure 13: Number of aggregate flow arrivals over 24 hours: simulated models vs. trace.

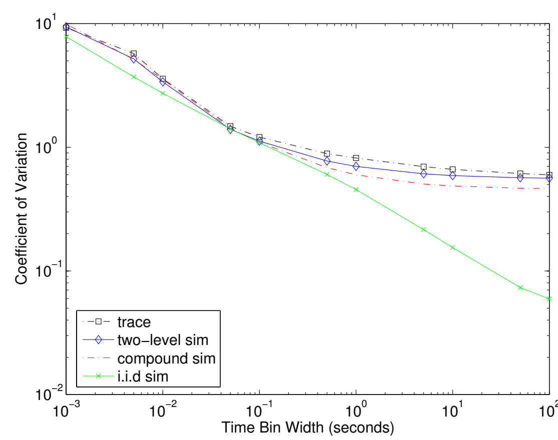

Figure 14: CoV of number of aggregate flow arrivals over different time intervals: simulated models vs. trace.

Finally, the CoV of the flow arrival count process over different time intervals is the subject of Figure 14. Our model matches very well the original trace throughout the different time scales outperforming the other two simulated models. The compound model exhibits equally good behaviour for small time intervals, but its deviation grows for higher time scales. For these scales, the deviation of the i.i.d model from the trace is even larger, making clear its inefficiency to capture the statistical structure of the trace.

6. AP-LEVEL MODELING

In this section, we investigate whether the two-level modeling approach for the traffic demand of the whole network can also be applied to individual APs. Intuitively, modeling

Table 2: Summary of our ap-level model (AP 222).

Modeled variable	Model	Parameters
Session arrival	Time-varying Poisson with rate $\lambda(t)$	Hourly rate: 1 (min), 928 (max), 11 (median)
Flow inter-arrival/session	Lognormal	$\mu = -1.6355$, $\sigma = 2.6286$
Flow number/session	BiPareto	$\alpha = 0.07, \beta = 1.75$, $c = 295.38, k = 1$
Flow Size	BiPareto	$\alpha = 0.00, \beta = 1.02$, $c = 15.56, k = 111$

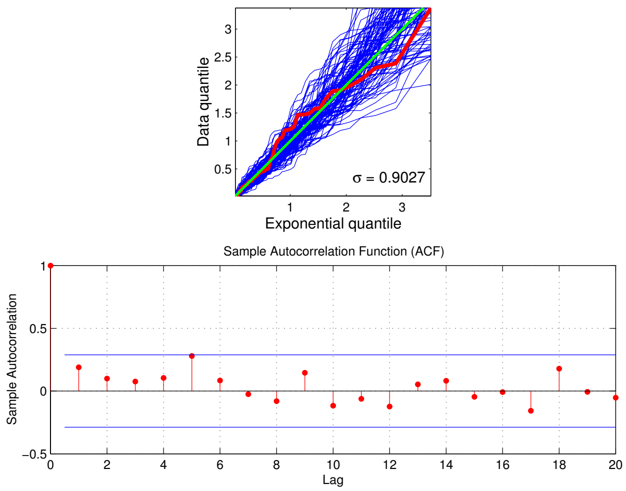

Figure 15: The R_{ij}s in AP 222 are independent and exponentially distributed. One randomly chosen hour is shown.

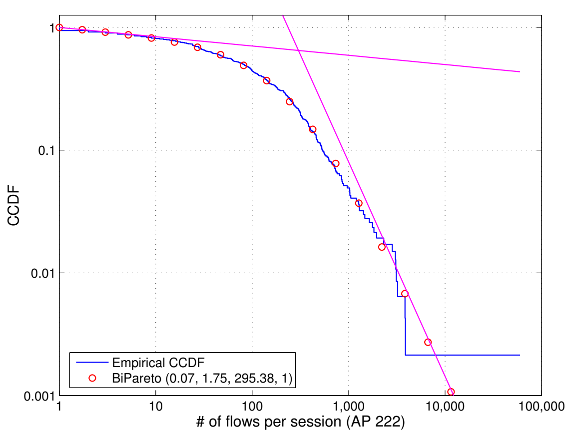

Figure 16: BiPareto model of number of flows per session in AP 222.

single APs is more difficult, since the reduction in the level of aggregation makes the data less well-behaved. However, we will demonstrate that the modeling insights from the system-wide modeling in Section 4 are also useful here, at least for selected hotspot APs of the wireless infrastructure. In the remainder of this section we focus on AP 222, one of the hotspots of the UNC wireless network. The statistical distributions derived for the traffic demand variables of AP 222 are summarized in Table 2.

Section 4.1 argues that the process of session arrivals at the entire wireless network can be described using a time-

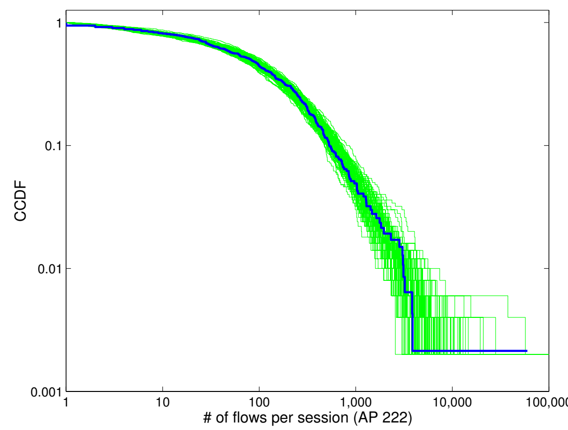

Figure 17: Simulation envelope for biPareto fit of flows per session in AP 222.

varying Poisson process. This is also the case for the process of session arrivals at AP 222. As in Section 4.1, we randomly select one hour during which there are more than ten session arrivals at AP 222, divide it into ten six-minute blocks and calculate the R_{ij}s according to (1). The top part of Figure 15 shows an exponential quantile plot of the R_{ij}s, which suggests that the exponential fit is clearly appropriate. The maximum likelihood estimate of the exponential parameter is 0.9027, which is very close to unity. The bottom plot of the figure illustrates the autocorrelations of the R_{ij}s up to 20 lags, from which one can tell that there is no much correlation among the R_{ij}s. We obtain similar results for all the hours during the 8-day trace, which have at least ten arrivals. The threshold of ten arrivals is chosen rather subjectively to ensure a large enough sample for the quantile plots.

The Poisson distributed session arrivals at AP 222 give rise to an interesting interpretation of the AP-preference function shown in Figure 3. It is well known that if a Poisson process is randomly partitioned into several point processes according to a set of fixed probabilities, the resulting point processes are still Poisson processes with rates proportional to the respective partition probabilities. In our study, the AP-preference probabilities may be viewed as the partition probabilities. As a result, the session arrival processes at separate APs should be approximately Poisson. This observation also supports the use of a simple algorithm for simulating session arrivals at specific APs. After simulating a certain number of sessions for the entire network, one can assign them to different APs using their corresponding AP-preference distribution.

When we consider a single AP, the number of flows per session can also be described with great accuracy using a biPareto distribution, as demonstrated in Figure 16. A biPareto simulation envelope is superimposed in Figure 17,

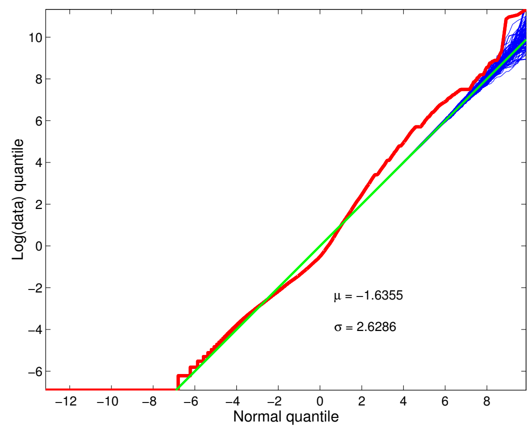

Figure 18: Flow inter-arrivals at AP 222 are well-modeled by a lognormal distribution.

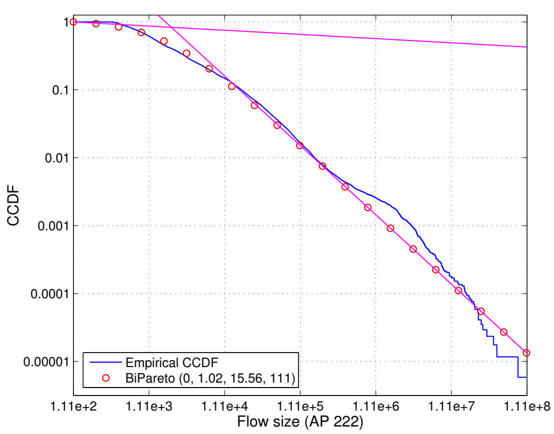

Figure 19: Model of flow size for AP 222.

which shows that the fit is clearly excellent, even for the values with the smallest probability located in the far part of the tail.

Next, we study the flow inter-arrivals within the sessions that started in AP 222, and the lognormal model proposed for the entire system remains applicable here. Figure 18 depicts the corresponding lognormal quantile plot. The two parameters are estimated to be -1.6355 and 2.6286 using maximum likelihood. Although the fit is worse than the one for the system-wide modeling, the quantile plot again follows the diagonal line closely, and the fit could still be useful.

Finally, a biPareto distribution yields an excellent fit for the size of flows that start from AP 222 in Figure 19.

7. RELATED WORK

Most traffic characterization studies focus on wired networks. Hierarchical approaches, looking at traffic variables above the packet-level, emerged in mid 90s. Flow-level traffic variables have been the subject of modeling in various studies, embracing almost all Internet protocols and applications, mainly TCP traffic [7, 6, 22, 9] but also multimedia streaming traffic [16]. The concept of session as a structure of the user activity was used in [23] for FTP traffic, as a synonymous of the FTP control connection. The term was used more explicitly later in Web traffic modeling. Both empirical [15, 24] and statistical [4, 18] modeling approaches have been used for the description of traffic at the two levels. A common feature of these studies is that the flow/session borders are heuristically defined by intervals of user inactivity. Our approach has been inspired by these studies, in particular from the approach of Nuzman *et al.* [18]. However, there are two main differences. Firstly, we relate the concept of session to the MAC-layer interaction of the user with the wireless network. Secondly, we do not adopt the scaling of in-session flow interarrival times according to the mean flow interarrival time, which is explicitly modeled in [18]. We rather fit in-session flow interarrival times directly to the trace data.

Fewer is the related work in wireless local area networks. The majority of the measurement studies [10, 3, 2] make high-level observations about traffic dynamics in both the temporal and spatial domain. Papadopouli *et al.* analyze the AP traffic patterns in various time scales and identify diurnal and weekly periodicites [20, 19], non-uniform distribution of workload across the wireless infrastructure [13], time-varying Poisson process client arrivals at APs, and building type dependencies [21]. To assess the impact of the wireless access on traffic characteristics, Hernández and Papadopouli [12] make a comparison of the wired and wireless traffic of the UNC campus with respect to flow-level traffic variables, such as connection duration, size and round-trip-time (RTT).

To the best of our knowledge, the only study that addresses the WLAN traffic modeling at higher detail is the one by Meng *et al.* [17]. It uses syslog and tcpdump traces from 31 APs in five buildings of the Dartmouth campus to model flow arrivals at 15 APs in one-hour intervals. They propose a Weibull distribution, and capture the non-stationarity of traffic in the variation of its scale parameter, which is estimated via Weibull regression. Furthermore, they model the flow size with a lognormal distribution. The authors find that a small percentage of the flows is roaming, *i.e.*, accessing data from more than one AP, and model the number of AP visits within an session with a geometrical distribution. They also observe strong similarity in the flow arrival processes at neighboring APs.

Contrary to [17], our work captures the non-stationarity of traffic workload at the session- rather than the flow-level via a time-varying Poisson process for session arrivals. We believe that this hierarchical approach provides better insight to the underlying causes of the *temporal* variations of the workload. Moreover, we use more data coming from a significantly larger number of APs, which allows us to see a significantly higher *spatial* variation of traffic load.

The modeling of traffic workload for each single AP over one-hour intervals, as proposed in [17], does not scale well. On the other hand, our AP-preference distribution approach is too coarse to model reliably the traffic demand spatial dynamics. In fact, selecting the appropriate scale for modeling the spatial characteristics of traffic workload is an open question that largely depends on the particular mechanism that needs to be analyzed. The AP-level can be problematic, since minor changes in the AP infrastructure, *e.g.*, addition of a new AP, may change significantly the workload distribution per AP. Higher levels of spatial aggregation, such as buildings or building types appear to be more appropriate in this context.

8. CONCLUSIONS

We present a hierarchical methodology for modeling the traffic demand in a campus wireless network. The two modeling levels are the wireless sessions and network flows. We investigate their statistical properties and inter-relations, deriving statistical distributions for a number of network-wide traffic demand variables, such as the session arrival rate, the flow number and their interarrivals within a session. The shift to sessions features two important advantages. Unlike visits to an AP, sessions can mask the network-related dependencies that are not important for a range of applications and system functions and exhibit nice statistical properties that make them amenable to modeling.

A standard challenge with measurement-based modeling is to find out how general is the validity of the derived models and up to what extent they can be reused. The validation of the models can be tried in different ways. In this paper, we use our models to generate synthetic traffic and compare it to the original trace with respect to traffic variables that have not been taken explicitly into account in our modeling approach. We find that the simulated traffic matches well the original trace. Interestingly, our modeling results also capture traffic demand characteristics in individual hotspot APs, implying that they can be used for modeling traffic workload over finer levels of spatial aggregation. As a further validation step, we are currently applying our modeling approach to measurement data obtained from UNC infrastructure during the last week of April 2006, i.e., a year after the tracing period of this paper. The first results suggest that the parametric distributions proposed in this paper hold for the new measurement data as well. A third, apparent, model validation step is the application of our models to measurement data collected from other infrastructures; however, as explained in Section 2, this is not always straightforward. We believe that better co-ordination of measurement efforts within the wireless networking research community will allow better reusability of measurement data and enable the coherent evaluation of models and tools.

Modeling the spatial dynamics of traffic load is challenging. In this paper, we look at this problem from two directions. We explore to what extent the findings of system-wide modeling pertain to lower levels of spatial aggregation (*i.e.*, hotspot AP). Furthermore, we derive the AP-preference distribution as a coarse abstraction of the spatial dynamics of the traffic load. We currently explore the spatial distribution of the network flows and sessions at various scales of spatial aggregation, such as the building, and building type. This information could be very beneficial in simulating different sizes of wireless networks and studying their spatial evolution.

A further refinement of our models will consider how the population size of wireless users relates to the process of session arrivals. Client dynamics are difficult to understand due to the wide range of behavior and pervasive non-stationarities. Some clients use the infrastructure only one or a few times and then disappear from the system, whereas others represent a more constant load. Understanding this part of the workload will make simulations more intuitive, since their input could be the number of clients and a parametric description of their access patterns.

Acknowledgment

This work was partially supported by the IBM Corporation under an IBM Faculty Award, a Marie Curie International Reintegration Grant, and the Hellenic Republic Ministry of Development, General Secretariat for Research and Technology, under a grant for "Cooperation with Scientific and Technological Institutions in non-European Countries, 2005".

9. REFERENCES

[1] Cisco Aironet AP specifications, http://www.cisco.com/en/us/products/hw/wireless.

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

[3] M. Balazinska and P. Castro. Characterizing mobility and network usage in a corporate wireless local-area network. In *Proc. of MobiSys*, San Francisco, CA, United States, May 2003.

[4] P. Barford and M. E. Crovella. Generating representative Web workloads for network and server performance evaluation. In *Proc. of ACM Sigmetrics*, pages 151–160, Madison, Wisconsin, June 1998.

[5] L. D. Brown, N. Gans, A. Mandelbaum, A. Sakov, H.Shen, S. Zeltyn, and L. Zhao. Statistical analysis of a telephone call center: a queueing-science perspective. *Journal of the American Statistical Association*, 100(469):36–50, March 2005.

[6] J. Cao, W. S. Cleveland, D. Lin, and D. X. Sun. On the nonstationarity of internet traffic. In *Proc. of ACM Sigmetrics*, pages 102–112, Cambridge, MA, United States, June 2001.

[7] W. S. Cleveland, D. Lin, and D. X. Sun. IP packet generation: statistical models for TCP start times based on connection-rate superposition. In *Proc. of ACM Sigmetrics*, pages 166–177, Santa Clara, CA, United States, June 2000.

[8] R. B. D'Agostino and M. A. Stephens. *Goodness-of-Fit Techniques*. Marcel Dekker, 1986.

[9] A. Feldmann. Characteristics of tcp connection arrivals. In *in Self-Similar Network Traffic And Performance Evaluation (K. Park and W. Willinger, eds.)*. John Wiley & Sons, 2000.

[10] T. Henderson, D. Kotz, and I. Abyzov. The changing usage of a mature campuswide wireless network. In *Proc. of ACM MobiCom*, Philadelphia, PA, United States, September 2004.

[11] F. Hernandez-Campos, M. Karaliopoulos, M. Papadopouli, and H. Shen. Spatio-temporal modeling of campus WLAN traffic demand. Technical Report 379, ICS-FORTH, Heraklion, Crete, Greece, June 2006.

[12] F. Hernandez-Campos and M. Papadopouli. Assessing the real impact of 802.11 WLANs: A large-scale comparison of wired and wireless traffic. In *14th IEEE Workshop on Local and Metropolitan Area Networks*, Chania, Crete, Greece, 2005.

[13] F. Hernandez-Campos and M. Papadopouli. A comparative measurement study of the workload of wireless access points in campus networks. In *16th Annual IEEE International Symposium on Personal*

Indoor and Mobile Radio Communications, Berlin, Germany, September 2005.

[14] P. Lewis and G. Shedler. Simulation of nonhomogeneous poisson process by thinning. *Naval Research Logistics Quarterly*, 26:403–413, 1979.

[15] B. A. Mah. An empirical model of HTTP network traffic. In *Proc. of IEEE Infocom*, pages 592–600, Kobe, Japan, April 1997.

[16] A. Mena and J. Heidemann. An empirical study of real audio traffic. In *Proc. of IEEE Infocom*, pages 101–110, Tel-Aviv, Israel, March 2000.

[17] X. G. Meng, S. H. Y. Wong, Y. Yuan, and S. Lu. Characterizing flows in large wireless data networks. In *Proc. of ACM MobiCom*, pages 174–186, New York, NY, United States, 2004.

[18] C. Nuzman, I. Saniee, W. Sweldens, and A. Weiss. A compound model for tcp connection arrivals for LAN and WAN applications. *Computer Networks*, 40(3):319–337, 2002.

[19] M. Papadopouli, E. Raftopoulos, and H. Shen. Evaluation of short-term traffic forecasting algorithms in wireless networks. In *2nd Conference on Next Generation Internet Design and Engineering*, Valencia, Spain, April 2006.

[20] M. Papadopouli, H. Shen, E. Raftopoulos, M. Ploumidis, and F. Hernandez-Campos. Short-term traffic forecasting in a campus-wide wireless network. In *16th Annual IEEE International Symposium on Personal Indoor and Mobile Radio Communications*, Berlin, Germany, 2005.

[21] M. Papadopouli, H. Shen, and M. Spanakis. Modeling client arrivals at access points in wireless campus-wide networks. In *14th IEEE Workshop on Local and Metropolitan Area Networks*, Chania, Crete, Greece, September 2005.

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

[23] V. Paxson and S. Floyd. Wide-area traffic: the failure of Poisson modeling. In *Proc. of ACM Sigcomm*, pages 257–268, London, United Kingdom, August 1994.

[24] F. D. Smith, F. Hernandez-Campos, K. Jeffay, and D. Ott. What TCP/IP protocol headers can tell us about the web. In *Proc. of ACM Sigmetrics*, pages 245–256, June 2001.

[25] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson. Self-similarity through high-variability: Statistical analysis of ethernet LAN traffic at the source level. *ACM CCR*, 25(4):100–113, October 1995.

APPENDIX

BiPareto distribution

The biPareto distribution is specified by four parameters (α, β, c and k), whose complementary cumulative distribution function (CCDF) is given by

$$\left(\frac{x}{k}\right)^{-\alpha}\left(\frac{x/k+c}{1+c}\right)^{\alpha-\beta}, \quad x \geq k.$$

$k > 0$ is the minimum value of a biPareto random variable, which is a scale parameter. The CCDF initially decays as a power law with exponent $\alpha > 0$. Then, in the vicinity of a breakpoint kc (with $c > 0$), the decay exponent gradually changes to $\beta > 0$.

Essentially, the biPareto distribution has two Pareto tails on both ends of the distribution. On a log-log plot, a CCDF of the form $x^{-\alpha}$ (*a Pareto tail*) would appear as a straight line with slope $-\alpha$. Thus, the log-log plot of a biPareto CCDF has two nearly linear regimes, with slopes $-\left(\frac{c}{1+c}\alpha + \frac{1}{1+c}\beta\right)$ and $-\beta$, respectively. This property of the distribution makes it a good choice for modeling the number of flows per session and flow sizes in Section 4. Its parameters can be estimated via maximum likelihood [18].

A Statistical Test for Time-varying Poisson Processes

In this section, we describe a test [5] for the null hypothesis that an arrival process is a time-varying Poisson process, with a slowly varying arrival rate.

To begin with, we break up the interval of a day into relatively short blocks of time. For convenience, blocks of equal length, L, are used, resulting in a total of I blocks; though this equality assumption can be relaxed. For the analysis in Section 4.1, L is chosen to be 0.1 hour.

Let T_{ij} denote the jth ordered arrival time in the ith block, $i = 1, \ldots, I$. Thus $T_{i1} \leq \ldots \leq T_{iJ(i)}$, where $J(i)$ denotes the total number of arrivals in the ith block. Define $T_{i0} = 0$ and

$$R_{ij} = (J(i) + 1 - j)\ln\left(\frac{L - T_{i,j-1}}{L - T_{ij}}\right), \quad j = 1, ..., J(i). \quad (1)$$

Under the null hypothesis that the arrival rate is constant within each time interval, the $\{R_{ij}\}$ will be independent standard exponential variables as we now discuss.

Let U_{ij} denote the jth (unordered) arrival time in the ith block. Then the assumed constant Poisson arrival rate within this block implies that, conditioning on $J(i)$, the unordered arrival times are independent and uniformly distributed between 0 and L. Denote $V_{ij} = \frac{L}{L-U_{ij}}$, and it follows that V_{ij} are independent standard exponential. Note that $T_{ij} = U_{i(j)}$, thus

$$V_{i(j)} = \ln\left(\frac{L}{L - U_{i(j)}}\right) = \ln\left(\frac{L}{L - T_{ij}}\right).$$

As one can see, $R_{ij} = (J(i) + 1 - j)\left(V_{i(j)} - V_{i(j-1)}\right)$. Then, the exponentiality of R_{ij} follows from the following well-known lemma.

Lemma: Suppose $X_1, \ldots, X_n$ are independent standard exponential, then $Y_i = (n-i+1)[X_{(i)} - X_{(i-1)}]$, $i = 2, \ldots, n$, are independent standard exponential.

Any customary test for the exponential distribution can then be applied to R_{ij} for testing the null hypothesis. For example, the familiar Kolmogorov-Smirnov test or Anderson-Darling test [8] could be used. However, as noted in [4], statistical significance tests are not very useful with large data sets, because they always tend to give insignificant results. Thus, we prefer to test the exponentiality hypothesis using a graphical tool, such as an exponential quantile plot with a simulation envelope as described in [11].