On scalable measurement-driven modeling of traffic demand in large WLANs

Merkouris Karaliopoulos [a] Maria Papadopouli [a,b,c] Elias Raftopoulos [b,c] Haipeng Shen [d]

a. Department of Computer Science, University of North Carolina, Chapel Hill, USA.
b. Institute of Computer Science, Foundation for Research and Technology - Hellas, Greece.
c. Department of Computer Science, University of Crete, Heraklion, Greece.
d. Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, USA.

Abstract— **Models of traffic demand are fundamental inputs to the design and engineering of data networks. In this paper we use real measurement data from a large wireless infrastructure to address this requirement in the context of large wireless networks. Our modeling effort focuses on capturing the demand variation in both the spatial and temporal domain in a way that scales well with the size of the wireless network. Traffic workload is modeled in terms of sessions and flows and buildings are viewed as the elementary entities of demand aggregation in the spatial dimension. We combine heuristics with clustering techniques to group buildings into clusters with similar traffic characteristics. Our modeling framework features elements that are reusable in the temporal and spatial dimensions. The parametric distributions for the session- and flow-related traffic variables successfully capture aggregate traffic demand in two different monitoring periods. Moreover, the same distributions can be used to model traffic demand at finer spatial scales, such as the single building or a group of buildings. Synthetic traffic is generated to compare our models with trace data and assess the trade-off between model scalability and reusability, on the one hand, and accuracy in capturing local-scale traffic dynamics on the other. Our main contribution is a novel methodology for traffic demand modeling in large wireless networks that features high flexibility in the exploitation of the spatial and temporal resolution available in data traces.**

I. Introduction

The modeling of traffic workload in large-scale wireless networks is the main focus of this paper. Although this task has been addressed in numerous research studies in the context of wired networks [1]–[4], there have been significantly fewer contributions for wireless networks.

One reason for this is that only recently have traces from large-scale wireless infrastructures with statistically significant network usage been made available. Furthermore, wireless network measurements are more complex than those in wired networks. Depending on how detailed view of the demand is required at the spatial dimension, e.g., at an access point (AP), APs co-located in a building or set of buildings, and the architecture of the network (single link-level subnet or multiple subnets), one needs to capture traffic at multiple physical locations. Since IEEE 802.11 MAC-layer frame sniffers are not commonly available, researchers often have to build custom equipment or resort to expensive commercial platforms for capturing the over-the-air traffic with the required level of detail. Finally, the transient characteristics of the radio propagation and user mobility make the analysis of the traffic-demand dynamics in both the spatial and temporal dimensions challenging. It comes as no surprise that the majority of the measurement studies, [5]–[7], make high-level observations about traffic dynamics in both the temporal and spatial domains without getting into the detail that modeling requires.

Arguments related to *scalability* and *reusability*, which are particularly desirable properties in modeling, complicate the problem further. Previous modeling studies have either attempted to model traffic demand over hourly intervals at the level of individual APs [8] or studied the problem at system-level deriving models for the aggregate network-wide traffic demand [9]. Clearly, both approaches have their strong and weak points. The second approach results in datasets that are amenable to statistical analysis and provides a concise summary of the traffic demand at system-level. However, it fails to capture the variation of this demand at finer spatial detail that may be required in the evaluation of system functions with focus on the AP-level (*e.g.*, load balancing). While working at the AP-level achieves that, it fails in other respects: the approach does not scale for large wireless infrastructures and data do not always lend to statistical analysis. Moreover, the modeling results are highly sensitive to the specific AP layout of a particular network and short-term variations of the radio propagation conditions.

This study has been motivated by the aforementioned challenges. We take advantage of the large wireless infrastructure of University of North Carolina (UNC) to obtain large amounts of measurement data. Our datasets provide considerable insights to the traffic load dynamics across the network and allow us to derive models of adequate detail for the traffic demand variation in space and time.

Our methodological choices attempt to strike a good trade-off between the two extreme approaches to traffic modeling that were outlined earlier, namely AP-level vs. network-level modeling. As in [9], we model traffic workload in terms of sessions and flows. Only now we look in more detail into the spatial dimension, using buildings as basic entities of traffic demand modeling. Major features of user activity, such as the traffic patterns they generate and their mobility within the wireless network, are studied at the building level. We then apply heuristics and more formal clustering techniques to group together buildings with similar traffic characteristics and achieve the scalability objective in our modeling.

Considerable effort is devoted to the validation of our modeling methodology. Synthesizing traffic after our models and comparing with the trace data, we assess the reusability of system-wide models to smaller spatial scales, the accuracy-scalability trade off and the possible contributions of clustering techniques to its resolution. Moreover, the availability of datasets from two different monitoring periods, spaced one year apart, lets us identify modeling elements that are time-persistent.

Our contributions are summarized in the following:

- A hierarchical framework for modeling traffic workload both system-wide and at finer spatial scales (*i.e.*, at building level and over groups of buildings). We find that the same set of parametric distributions describe our session- and flow-related traffic variables at various spatial scales and over two different monitoring periods.
- A novel methodology for scalable modeling of the spatial variation of traffic demand in large wireless networks drawing on heuristics and statistical clustering techniques.
- Validation of our modeling approach assessing the model accuracy and scalability. Our results suggest that the two approaches, heuristics and clustering, are complementary in that they result in clusters with high purity in different traffic variables.
- A set of analysis tools that have been made publicly available to the research community to enable further comparative studies [10].

The next section briefly describes the UNC wireless network infrastructure and the collected traces. Section III outlines the principal building blocks of our modeling approach, while Section IV focuses on the spatial variation of the traffic variables we model at various level of spatial aggregation. We apply clustering techniques to our problem in Section V and assess the modeling alternatives coming out of it in Section VI. Related work is reviewed in Section VII, while Section VIII concludes this report summarizing our main findings.

II. Data collection and processing

Two types of data are used in this study, packet header traces and Simple Network Management Protocol (SNMP) data, drawn from the wireless network of the UNC campus at Chapel Hill. We analyze datasets coming from two separate eight-day long monitoring periods; the first dataset corresponds to the period April 13-20, 2005, whereas the second one covers the interval Apr 28-May 5, 2006.

The UNC campus wireless network comprised 488 APs by April 2005 and 741 APs one year later. Almost all of them belong to the Cisco Aironet series [11] and they are standalone APs according to the terminology in [12][1]. The network APs are spread over more than 220 in-campus buildings, including student residence halls, academic buildings, sport halls, and libraries, and a few off-campus administrative offices, providing wireless access to 26,000 students, 3,000 faculty and 9,000 staff members.

SNMP data are collected from all the network APs with a period of five minutes. We implemented a custom SNMP-polling system relying on a non-blocking SNMP library. APs are polled independently, so that delays incurring during the processing of SNMP polls by the slower APs do not affect the other APs. The collection of SNMP data is a 24/7 process, which has been running almost without interruption since September 2004.

Packet header traces are collected with a high-precision monitoring card (Endace 4.3GE). The card was installed in a high-end FreeBSD server and captured all packets traversing the link between UNC and the Internet in both directions. The monitoring period was 178.2 hours in April 2005 and 192 hours in April 2006, yielding 175GB and 365GB of packet headers respectively. The sharp increase in the collected amount of packet headers is primarily due to the significant growth of the network infrastructure between the tracing periods. Our measurement data are summarized in Table I.

A dedicated set of IP addresses is reserved in the UNC network for WLAN clients. They are dynamically assigned via DHCP an IP address and they maintain it as they roam within the network, *i.e.*, the campus backbone network behaves like a single link-layer domain. Filtering with the respective address prefixes, we can extract the wireless portion of traffic out of the full UNC campus traffic, wired and wireless, monitored in the link connecting the UNC campus with the Internet. More significantly, we can directly correlate the SNMP data

[1] In summer 2006, these standalone APs were replaced by "thin" APs and Wireless Network Controllers of the same brand

drawn from the APs with the packet header traces and infer sessions and connections, which are central to our modeling approach as explained in Section III.

III. Modeling methodology

A. *Wireless sessions and network flows*

Starting point for the work presented here are the results presented in [9] for the aggregate network traffic demand. We adopt a hierarchical modeling approach that organizes the client activity into two levels, the wireless session and the network flow. The wireless session delineates the interval that the user is connected to the infrastructure and active in producing traffic. It 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. In our approach, sessions account for the traffic non-stationarity in time and are modeled by a time-varying Poisson process. On the other hand, network flows, such as TCP connections and UDP conversations, are well-separated collections of packets between a pair of Internet hosts, *i.e.*, packets that share the same transport-layer "5-tuple". The well-established advantage of flow-level modeling is its higher independence from the specific network topology and measurement conditions when compared with packet-level modeling [13]. The flow-related session attributes we model are the in-session number of flows, in-session flow interarrivals and flow sizes. In the floowing two paragraphs, we describe how we derive these attributes out of out measurement data.

1) Wireless session duration inference: It is possible to infer the duration of wireless sessions using either the Syslog or the SNMP measurement data. Each method has its own inherent advantages and weaknesses, effectively posing hard constraints to the accuracy of the modeling approach.

Syslog messages are event-based. It is thus feasible to know exactly when a WLAN client was (re)associated/disassociated to/from the network. Nevertheless, the problem is that the client is not producing traffic (active) throughout his association with the network. Taking a more detailed look into at the measurement data, we found that there is a significant number of cases, where the client remains connected to the network, although there is no use of the wireless device (usually laptop). In these cases, the disassociation of the client only takes place after a given interval of client inactivity, upon expiration of a protocol timer. As a result, relying on Syslog messages to infer the end of a client session provides a positively biased estimate of the session duration, as shown in Figure 1b. Similar concerns, although at smaller extent, relate to the inference of the session start.

On the other hand, SNMP pollings allow to figure out when a client is actually producing traffic and when it stays idle. Therefore, the estimates for the session duration do not suffer from the bias related to Syslog messages. However, the disadvantage of SNMP data is their resolution, which is upper bounded by the SNMP polling period. Whereas the information drawn from SNMP tables allows to infer precisely the start of the session, the end of the user activity can be known only with a precision in the order of the SNMP polling period (see Figure 1c). Even worse, the SNMP polls fail to report on clients that associated to and disassociated from one or more APs inbetween two successive SNMP polling instants, as shown in Figure 1d.

Despite their lower resolution, SNMP data result in tighter estimates for the client session duration. Therefore, they were taken as the basis for inferring the start and end of client sessions and the derivation of the modeled flow-related attributes, as it will be discussed in the next paragraph. It is left as future work to investigate the additional accuracy we can gain by considering jointly the Syslog and the SNMP data in inferring the client session durations.

2) Correlation with packet header data: Irrespective of the way the wireless session duration is estimated, we need to derive the number of flows and the flow interarrival durations within each session and the flow sizes. This task is simpler. The SNMP tables provide the IP address assigned to a client during his session. Remember that this address remains the same irrespective of where the client moves within the UNC campus WLAN, *i.e.*, the backbone network behaves as a single link-layer domain. When sessions exceed the DHCP lease period (1 hour), the same address is reassigned to the WLAN client upon the DHCP lease renewal. Two or more IP addresses for a client during a given session

TABLE I

Analyzed measurement datasets from UNC campus

Dataset	Monitoring period	Number of APs	Number of WLAN clients seen	Size
Packet headers 05 SNMP pollings 05	Apr 13-20 2005	488	9777	175GB 320MB(compressed)
Packet headers 06 SNMP pollings 06	Apr 28 - May 5 2006	741	12484	365GB 657MB(compressed)

were only noted in a small percentage of sessions.

The start and end times of a network flow (TCP connection or UDP conversation) can be extracted from the packet header traces. In this case, inaccuracy may be due to the loss of packets, in particular of those signalling the start/end of a flow (*e.g.*, TCP SYN, SYN/ACK, FIN packets). When this happens, the start(end) of a flow is assumed to coincide with the timestamp of the first(last) packet seen on the wire for the specific flow.

With the four time instants at hand (start and end times of wireless sessions and flows), it is then straightforward to derive the number of flows carried out within a session, the interarrivals between flows within a given session and the flowsizes, as outlined in Figure 2.

3) Network-wide modeling distributions: Notably, we found out that the same statistical distributions, though with different parameter sets, model our traffic variables in both the 2005 and 2006 monitoring periods. The in-session number of flows and the flow sizes are well modeled by the BiPareto distribution, whereas the Lognormal distribution is the best fit for in-session flow interarrivals out of a set of common distributions, including Weibull, Gamma, and Pareto. A time-varying Poisson process with constant rate over intervals of an hour captures the non-stationarity of session arrivals. The results for the 2005 monitoring period are detailed in [9], whereas the fitted distributions for the 2006 dataset are shown in Figure 3. In the same figure, we show sample results of the statistical test for the time-varying Poisson process of session arrivals within intervals of an hour [14]. The same results are obtained for all one-hour intervals in our 192-hour long 2006 packet header trace. Table II summarizes the distributions and their parameters for the two periods.

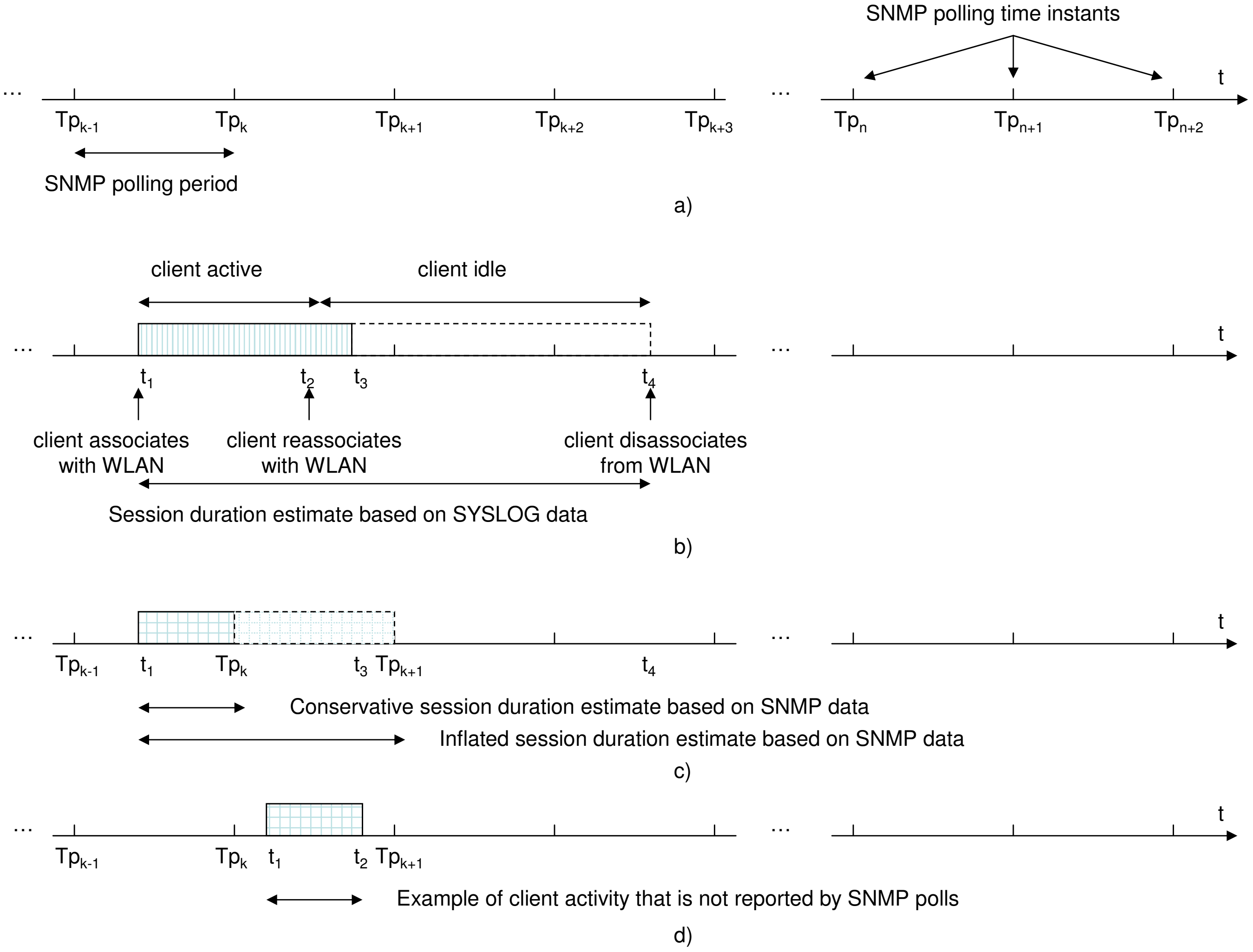

SNMP polling time instants
t
...
Tp_{k-1}
Tp_k
Tp_{k+1}
Tp_{k+2}
Tp_{k+3}
Tp_n
Tp_{n+1}
Tp_{n+2}
SNMP polling period
a)
client active
client idle
t_1
t_2
t_3
t_4
client associates with WLAN
client reassociates with WLAN
client disassociates from WLAN
Session duration estimate based on SYSLOG data
b)
Conservative session duration estimate based on SNMP data
Inflated session duration estimate based on SNMP data
c)
Example of client activity that is not reported by SNMP polls
d)

Fig. 1. Estimating the wireless session duration out of the SNMP and the SYSLOG data.

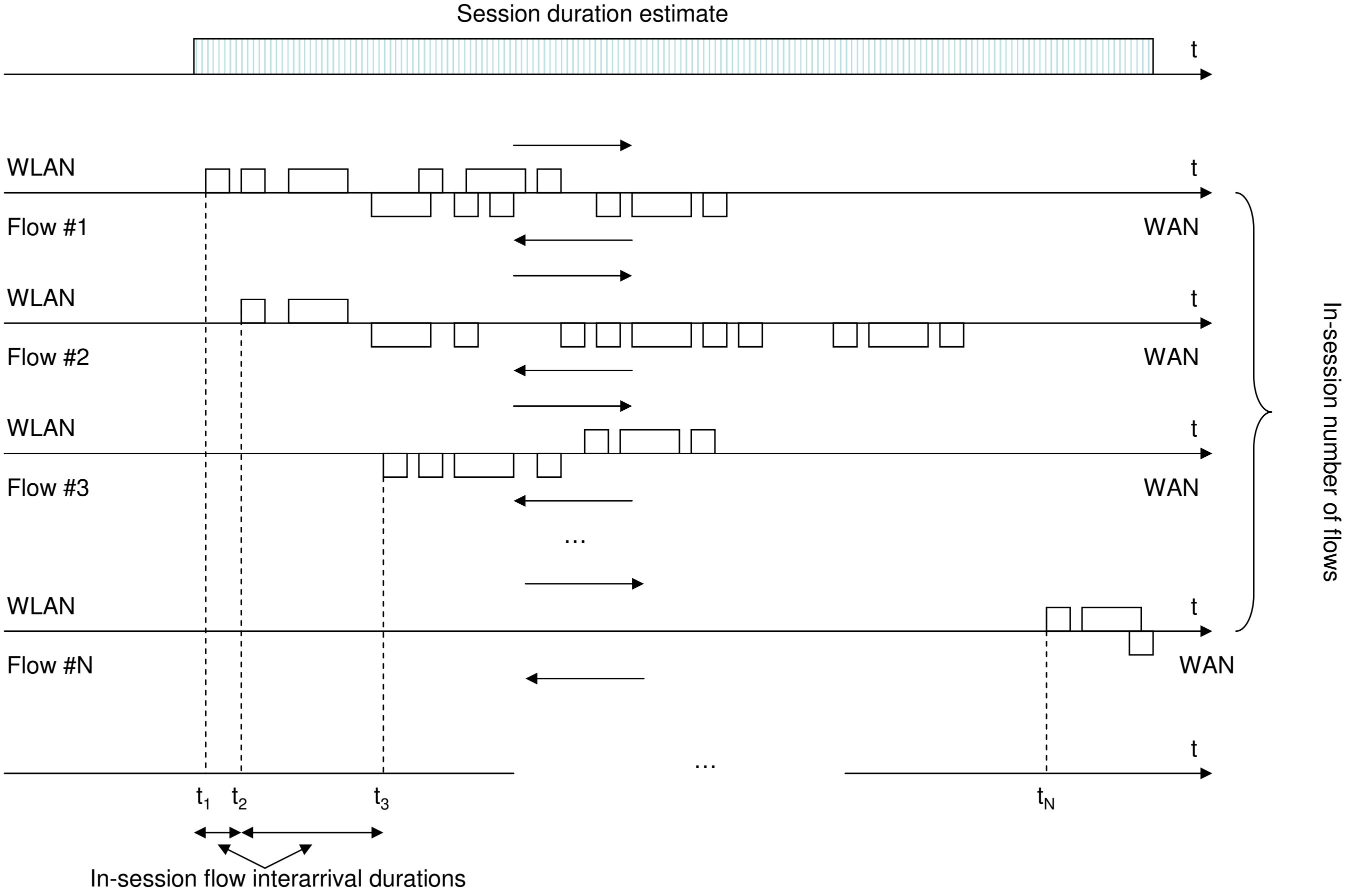

Session duration estimate
t
WLAN
Flow #1
WAN
WLAN
Flow #2
WAN
WLAN
Flow #3
WAN
...
WLAN
Flow #N
WAN
In-session number of flows
t
t$_1$
t$_2$
t$_3$
t$_N$
...
In-session flow interarrival durations

Fig. 2. Relating network flows to wireless sessions.

TABLE II

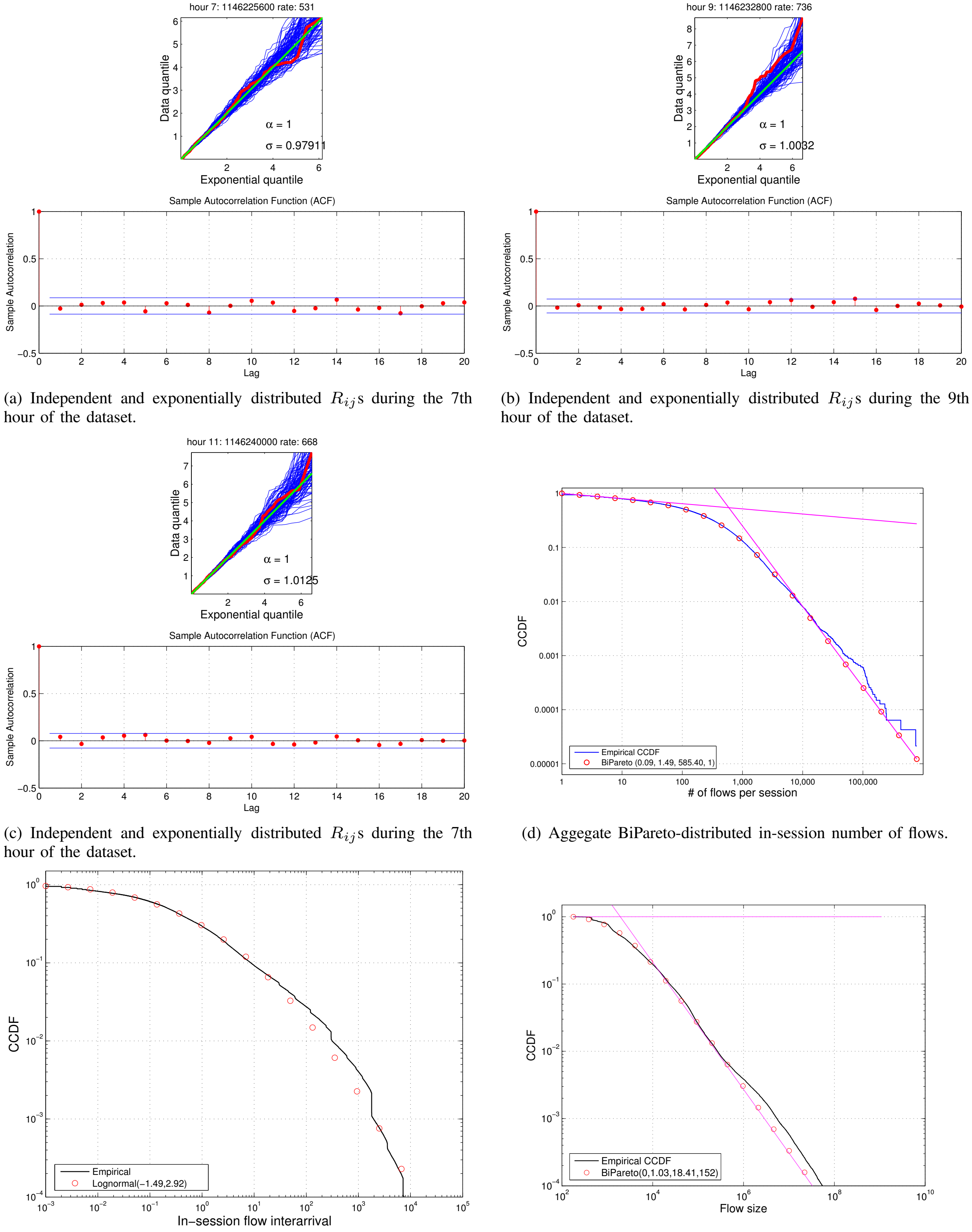
SUMMARY OF MODELS FOR NETWORK-WIDE TRAFFIC DEMAND VARIABLES.

Modeled variable	Model	Probability Density Function (PDF)	Parameters 2005	Parameters 2006
Session arrival	Time-varying Poisson($\lambda(t)$)	N: # of sessions between t_1 and t_2 $\lambda = \int\limits_{t_1}^{t_2} \lambda(t)dt,\ Pr(N = n) = \frac{e^{-\lambda}\lambda^n}{n!},\ n = 0, 1, \ldots$	Hourly rate: 44 (min), 1132 (max), 294 (med.)	Hourly rate: 75 (min), 1171 (max), 460 (med.)
Flow interarrival/session	Lognormal	$p(x) = \frac{1}{\sqrt{2\pi}x\sigma} \exp\left[-\frac{(\ln x - \mu)2}{2\sigma 2}\right]$	$\mu = -1.37,\ \sigma = 2.79$	$\mu = -1.49,\ \sigma = 2.92$
Flow number/session	BiPareto	$p(x) = k^{\beta}(1 + c)^{\beta-\alpha}x^{-(\alpha+1)}(x + kc)^{\alpha-\beta-1}$ $(\beta x + \alpha kc)\ ,\ x \geq k$	$\alpha = 0.06, \beta = 1.72,$ $c = 284.79, k = 1$	$\alpha = 0.09, \beta = 1.49,$ $c = 585.4, k = 1$
Flow size	BiPareto	Same as above	$\alpha = 0.00, \beta = 0.91,$ $c = 5.20, k = 179$	$\alpha = 0.00, \beta = 1.03,$ $c = 18.41, k = 152$

B. Buildings rather than APs

Whereas modeling traffic demand at system-level gives a good insight to the user activity patterns and forms a valuable input for the network design and dimensioning, many system functions work at smaller spatial scales. For example, load balancing algorithms usually consider the traffic load of a set of APs that are in close proximity when making their decisions. Evaluation of these functions requires traffic demand models of finer detail in the spatial dimension.

There are more than one ways to study and capture the spatial traffic demand variation in our wireless network. In fact, in that same work [9], two possibilities are described. The one offering the most detail is the separate modeling of each network AP. The same set of traffic variables that are explicitly modeled for the network-wide traffic demand, may be modeled for individual APs. This is also the approach followed in [8], although the actual modeling decisions are different (see discussion in Section VII). Another approach is to start from the



(a) Independent and exponentially distributed R_{ij}s during the 7th hour of the dataset.

(b) Independent and exponentially distributed R_{ij}s during the 9th hour of the dataset.

(c) Independent and exponentially distributed R_{ij}s during the 7th hour of the dataset.

(d) Aggegate BiPareto-distributed in-session number of flows.

(e) Aggregate Lognormal-distributed in-session duration of flow interarrivals.

(f) Aggregate BiPareto-distributed flow sizes.

Fig. 3. Network-wide flow-related attributes.

single set of statistical distributions in Table II and rely on weighting functions for capturing the spatial variation. The AP-preference distribution in [9] is one such example. The distribution defines the probability with which a session of the aggregate session arrival stream is initiated at a network AP.

One of the main advantages of working at the system-level is that there are statistically significant data for all modeled variables. This is not always the case with individual APs; in fact, only a limited number of hot-spot APs are amenable to modeling. A second major concern with AP-level modeling is scalability. The number of distributions that have to be derived and simulated grows linearly with the number of APs, which is not desirable when studying large-scale wireless infrastructures.

In our approach, we work with buildings. We view buildings as more reliable entities for modeling the spatial variation of traffic demand. In fact, we could draw an analogy between flows-packets and buildings-APs. Much as packet-level dynamics are subject to network topology and instantaneous conditions, AP-level user activity is sensitive to radio propagation dynamics and environmental setting. One good example is the ”ping-pong” effect, where a stationary user may be alternately associated with two, or even more, APs due to short-term radio signal propagation variations.

The notable advantage of working with buildings is that many of our findings for the aggregate network traffic demand also hold for the per building traffic demand. Session arrivals, for example, can still be modeled by time-varying Poisson processes. Figure 4 shows the exponential quantile plots and autocorrelation functions for the variables R_{ij}s, which are functions of the session arrival time series in several campus buildings.

As explained in [9], under the null hypothesis that the arrival rate is constant within each time interval (here, an hour), the $\{R_{ij}\}$ will be independent standard exponential variables. The maximum likelihood estimate of the exponential parameter are plotted along with the exponential quantile plots and are reasonably close to unity. The bottom plots of the figures are the autocorrelations of the R_{ij}s up to 20 lags. The sample autocorrelations are always within the confidence intervals, suggesting that the R_{ij}s do not exhibit any significant correlations. We got similar results when repeating the same analysis for other buildings.

In our network there are approximately 250 buildings, which can be grouped according to their main/exclusive usage into nine main categories. Table III lists the main building categories and their number in the UNC campus. In the following, one of the questions we attempt to answer is what level of modeling the traffic demand variation in the spatial dimension yields the best trade-off between modeling efficiency and scalability. In generally, for each traffic variable listed in Table II, the spatial detail of modeling could be the building, building type, or in the extreme case the network as a whole. Besides these intuitive ways to aggregate data, we apply clustering techniques for grouping buildings with similar traffic characteristics. As it will be shown in Sections V and VI, clustering allows us to better address the tradeoff between scalability and model accuracy.

IV. Spatial characteristics of traffic demand

The type of building, the population of clients that access the network, the patterns of usage, and the environment are a non-exhaustive list of factors that contribute to the spatial and temporal variation of traffic demand. In this section, we show how the modeled traffic variables of Table II vary across various time (hour, day, week) and spatial scales (building, building-type).

A. *Variation of session-arrival rate within day/week*

Figure 5 plots the hourly session arrivals over the whole trace duration (192 hours) for some representative campus buildings. Although the absolute numbers of session arrivals and their exact variation are specific to each building, these profiles exhibit clear patterns that are, to a large extent, intuitive and closely related to the building type and usage. For example:

- Administrative and business buildings show strongly similar daily and weekly patterns in their profiles. The activity window is quite narrow during weekdays (6-8 hours long), in agreement with the working hours, whereas the activity during weekend is almost zero.
- Residential buildings show distinctly different patterns. The number of session arrivals is more uniformly distributed across the week and hours within the day. The activity is also significant during the evening hours, often resulting in a daily or weekly peak.
- Academic buildings lie somewhere in between these two patterns. The daily window of activity is clearly broader than administrative and business buildings, since they host WLAN clients for longer time intervals during the day. Weekends see fewer session arrivals and shorter windows of activity when compared with residential buildings, but traffic is non-negligible.

B. *Variation of session-level flow-related variables*

The variation of traffic demand is also evident in the session-level variables we model. One way to see this variation is to draw their empirical distribution functions at the building-type level, as shown in Figure 6. Figure 6(a) shows the broad variation of the per building-type

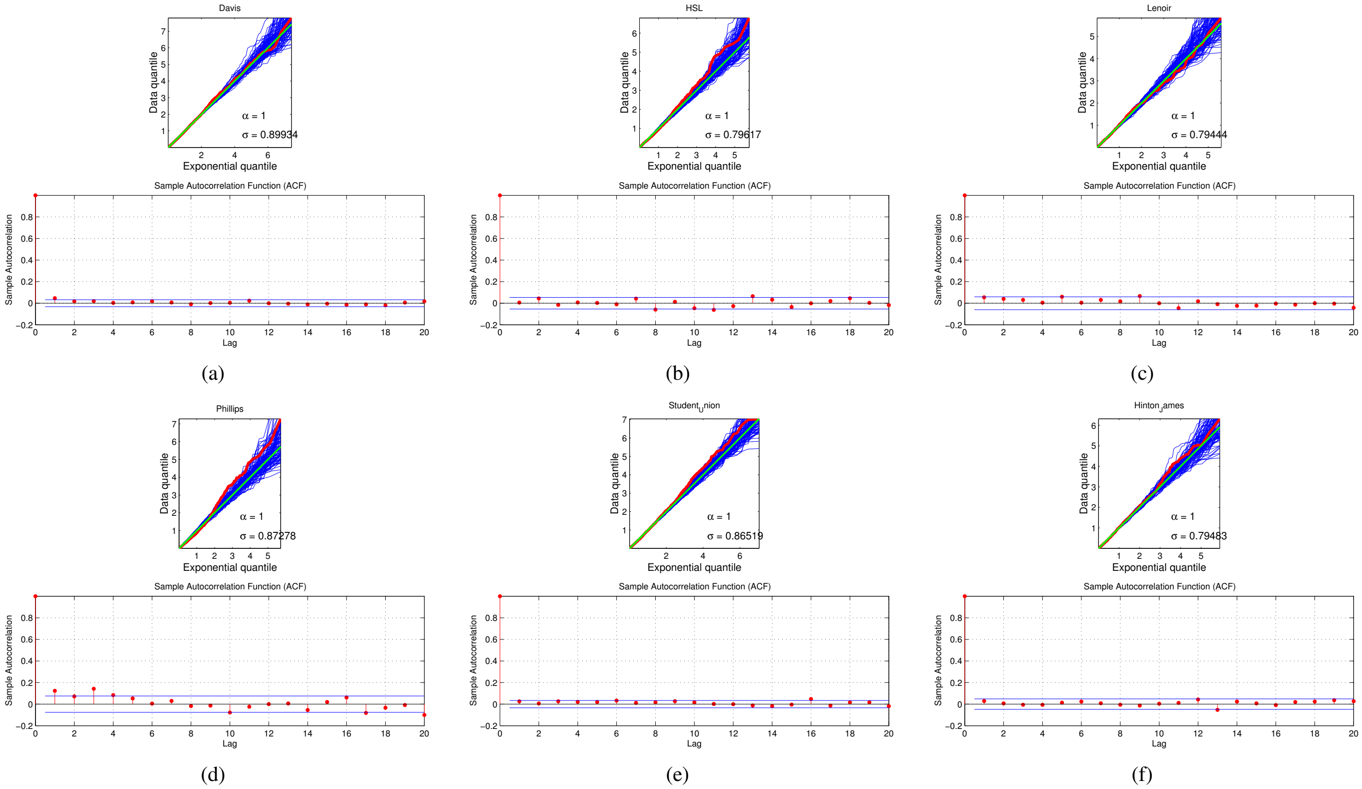


Fig. 4. QQplots and autocorrelation function for the R_{ij}s for different campus buildings. R_{ij}s are independent and exponentially distributed.

TABLE III
SUMMARY OF UNC CAMPUS BUILDING TYPES.

Building type	Academic	Administrative	Athletic	Business	Clinical	Library	Residential	Social
Number	51	25	17	8	18	4	117	10

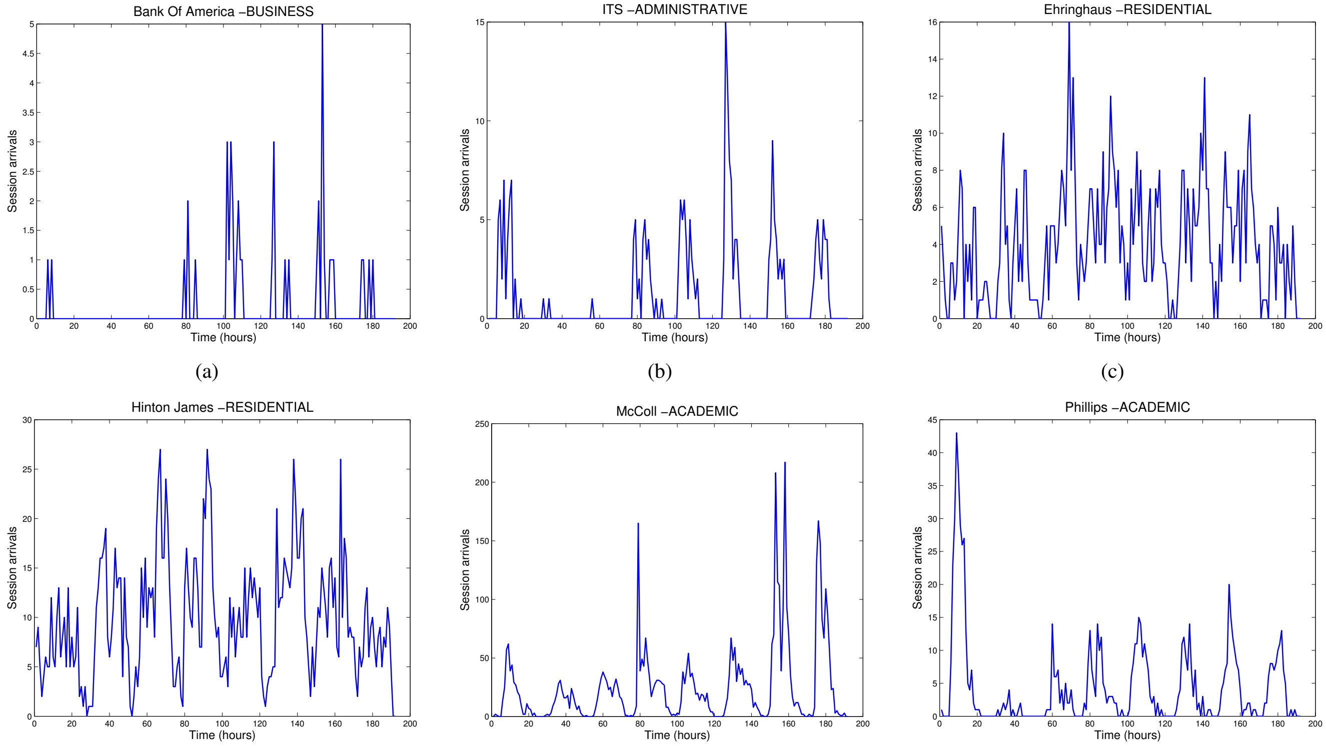
distribution tails of the in-session number of flows. The number of flows related to residential buildings sessions has a strikingly heavier tail, largely related to the more active Web browsing behavior of residential users. The plots also suggest that the BiPareto distribution can be applied for modeling the per building-type in-session number of flows. Table IV lists the parameter sets for the different building types.

More similar are mean flow sizes across different building types. Figure 6(b) suggests that some building types cluster together, such as Library, Residential and Academic, Administrative, Athletic, Social with flow sizes in Clinical buildings having a more distinct behavior, yet closer to the 2nd group of building types.

The behavior of flow interarrivals across different building types is captured in Figure 6(c). Again, the plots of mean in-session flow interarrivals suggest that the variables could be potentially modeled by the same type of distribution for all building types, though with different parameters.

Less wide is the differentiation of client sessions with respect to mobility, at least when this is viewed at the building type level. *Building-roaming* sessions, during which a WLAN client visits more than one building, account for less than 10% of the overall sessions. Figure 6(d) plots the per building type *building path length distribution*, expressing the probability that a session initiated at buildings of a certain type visits a given number of buildings. The plot clearly shows that building-roaming flows are a small percentage of the overall client sessions and there is little dependence on what kind of building a session is initiated.

Overall, the plots in Figure 6 show clearly that the modeled traffic variables exhibit strong variation in the spatial dimension. Although the building type is an intuitive, heuristical basis for grouping buildings, it is not the best one. In fact, in the section that follows, we use clustering techniques to come up with alternative groupings of buildings of higher utility for our modeling task.



(a) (b) (c)

(d) (e) (f)

Fig. 5. Hourly session arrival rates for representative UNC campus buildings over the Apr-May 2006 monitoring period.

TABLE IV

BIPARETO DISTRIBUTION PARAMETERS FOR THE IN-SESSION NUMBER OF FLOWS IN DIFFERENT BUILDING TYPES

Building type	Academic	Administrative	Athletic	Business	Clinical	Library	Residential	Social
BiPareto Parameters (α, β, c, k)	(0.11, 2.15, 702.99, 1)	(0.15, 1.73, 523.72, 1)	(0.15, 1.65, 1033.84, 1)	(0.16, 2.39, 1008.76, 1)	(0.13, 2.6, 819.5, 1)	(0.06, 2.33, 862.24, 1)	(0.08, 1.34, 961.71, 1)	(0.09, 2.24, 571.21, 1)

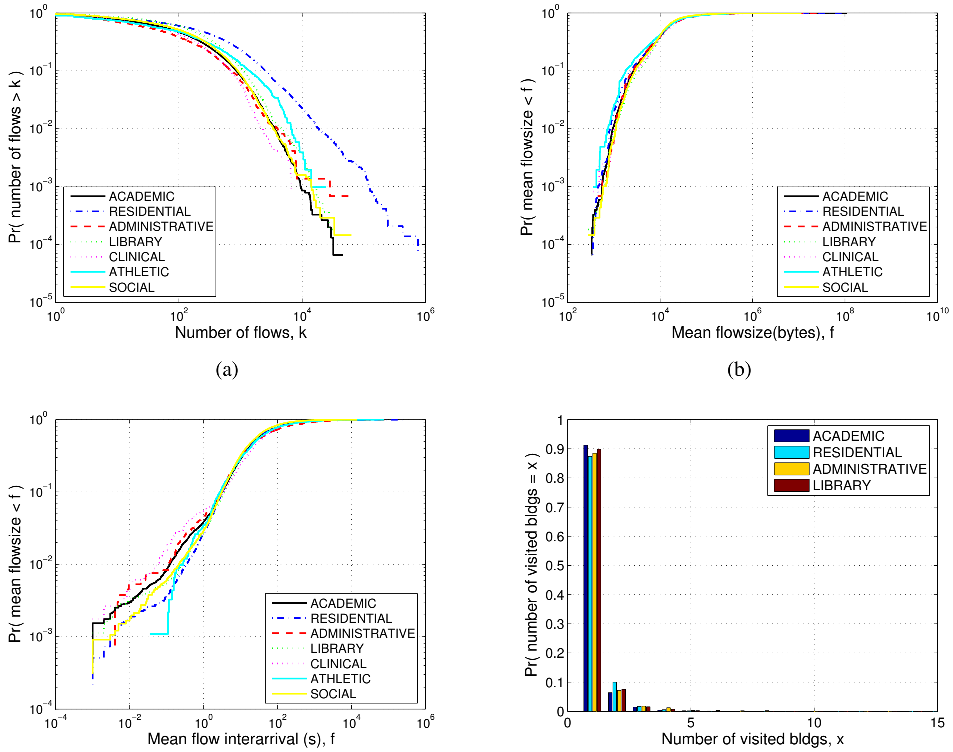
V. ENHANCING MODELING SCALABILITY WITH CLUSTERING

Our hierarchical modeling framework evolves around individual buildings at the finest and the entire system at the coarsest detail. As mentioned in the introduction, both approaches have weaknesses. We address them by enhancing this framework with an intermediate level of detail, namely *clusters* of buildings that exhibit similar behavior with respect to the traffic variables we model.

A. Clustering with respect to session arrival rate

Sessions are the main modeled entities in our two-level model and the ones where the traffic non-stationarity is captured via the time-varying Poisson processes, as shown in Section III-B Hence, we apply clustering techniques at the session level with the aim to come up with groups of buildings featuring similar variation of session arrivals in time. We work with the 2006 trace resulting in a time series of 192 hourly session arrival rates for each building. The time series are the *features* or *attributes* of the data matrix X input to clustering. The matrix has 250 rows, one for each campus building.

The ultimate aim of our building clustering is to use the same cluster-level hourly session arrival rate time series, hereafter called *cluster profile* or *signature*, for a group of buildings instead of a separate time series for each building, the *building profile*. Therefore, our clustering needs to take into account the size displacement between different building profiles. This requirement cannot be satisfied if we consider heuristic ways to group buildings, *e.g.*, the building type as defined in Section III, which result in building groups with high similarity in shape but large size displacements. To get clusters of buildings with the desired properties we combine clustering with dimension reduction techniques. Whereas the clustering algorithm is the same in all cases, we consider three alternatives for reducing the dimensionality of our



(a)

(b)

(c)

(d)

Fig. 6. Behavior of session attributes across different types of campus buildings.

dataset and bring out the size element. They rely on *Principal Component Analysis* (PCA) and *Singular Value Decomposition* (SVD).

1) Data dimensionality reduction via PCA and SVD: In the first method, centering is applied to the data matrix by subtracting the column means of X, which correspond to the average building session arrival rates. We then perform PCA [15]. The aim is to reduce the dimensionality of the dataset whilst preserving most of the variation in data in the first *few* Principal Components (PCs). To decide how many PCs to extract, we employ the *scree plot*, which plots the percentage of variance in data that is explained by PC i.

The other two alternatives do not apply any centering on the data matrix. In the first approach, we take the original data matrix X and standardize the time series for each building by dividing over the scale factor, calculated as the square-root of the sum of squares of session arrivals over the 192 hours. We then apply SVD to the standardized matrix to obtain the same number of left PCs as the one that came out of the PCA-based method described earlier. The scale factor vector is then used as an additional dimension, in addition to the PCs that came out of SVD, to be input to the actual clustering algorithm. The second alternative is similar to the first one, only now the additional dimension is the vector of average numbers of session arrivals over the whole tracing period for the 250 buildings.

In the rest of the section we present in more detail the clustering results for the PCA-based clustering. We then give the results that came out of the two SVD-based clustering alternatives and compare them on the basis of validity indices assessing the compactness of clusters and their degree of separation.

2) Building Clustering: The first thing to do before proceeding with the actual clustering is the determination of the number of PCs to use. Figure 7 shows the scree plot for the first seven PCs. The knee is located at PC3 indicating that the first three components can capture most of the variability in our dataset (approximately 90%). The plots of the first three loading vectors in Figure 8 suggest that PC1 is highly correlated with the mean session arrival rate, PC2 captures the difference between day (6am-7pm) and night, whereas PC3 expresses the difference between the first 12 hours of day (12pm - 11am) and the last 12 hours of day.

Clustering is subsequently carried out on the *projected space*, *i.e.*, the space spanned by the orthonormal PC1-PC3. We employ agglomerative hierarchical clustering. Dissimilarity between buildings and clusters is measured by the standardized Euclidean distance and the unweighted pair-group method using arithmetic averages

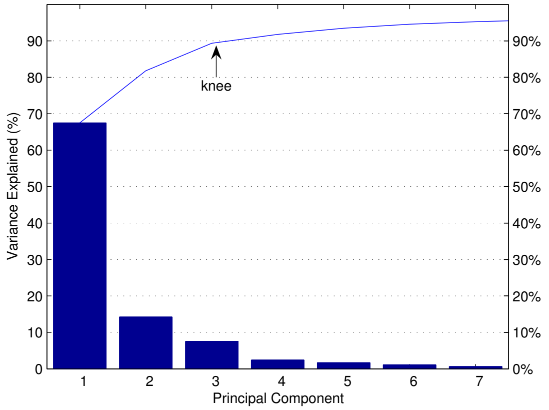


Fig. 7. Percentage of variance explained by principle component

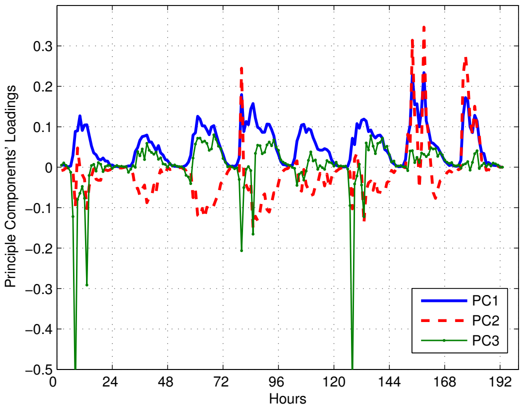


Fig. 8. Principle Components.

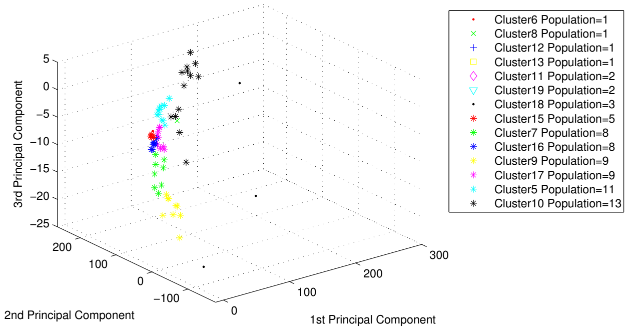


Fig. 9. 3-D Plot of PC1-PC3.

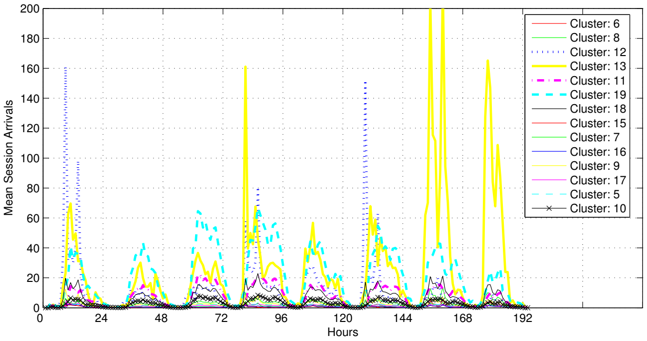


Fig. 10. Cluster Profiles.

(UPGMA), respectively. They are both popular choices for clustering tasks [16]. The final segmentation of buildings is the one that maximizes cluster compactness and separation, as measured by the inconsistency coefficient [17].

An additional processing step we take *after* clustering is to exclude from subsequent analysis buildings with few session arrivals. Those buildings actually group together in the same cluster. To assist visualization of results and since the major challenges to system engineering come from the heavily loaded buildings, we use heuristics to filter them out. After trial and error, the rule we set for filtering is to exclude those clusters, in which all buildings have mean session arrival less than 50, and the average of their maximum hourly session arrival rate does not exceed a threshold equal to 5. The remaining 74 buildings are shown in Figure 9.

Figure 9 plots the PC1-PC3 using different colors and symbols to present the various clusters obtained by the hierarchical clustering. To characterize each cluster, we calculate the cluster centroid in the projected space. To capture the main behavioral and statistical characteristics of each cluster, we then back-transform its centroid to get the "signature" session arrival rate series that corresponds to the specific centroid. This signature then works as a profile for that particular cluster. Figure 10 plots the signature profiles for the obtained clusters. For example, clusters 5 and 10, which are mostly populated by residential buildings, exhibit a high arrival rate during afternoons and a similar pattern during weekends; on the other hand, clusters 9 and 16 consist of academic buildings, which have a strong session arrival rate peak during the mornings on weekdays and a distinct drop during weekends.

3) Cluster validation: To validate the clustering result, we can use some internal criteria to measure how well a clustering fits the geometric structure of the data with no reference to information known a priori, such as the *silhouette index* [18]. A silhouette index value close to unity implies that the corresponding building is assigned to an appropriate cluster. An index value close to zero suggests that the building could also be assigned to the nearest neighboring cluster, i.e. such a building lies equally between both clusters. A building should be considered misclassified when the index is close to -1. Moreover, for any given partition, we can define the global silhouette index, *GSu*. *GSu* equals to the average silhouette index over all clusters and can be used as an effective validity index for a given clustering. The plot of the silhouette indices for the 74 buildings that

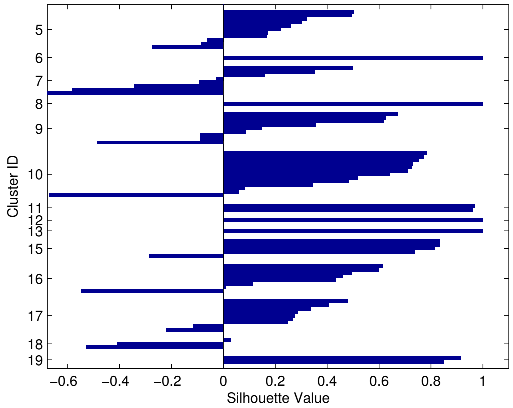


Fig. 11. Silhouette cluster validation index for the 192h feature-set.

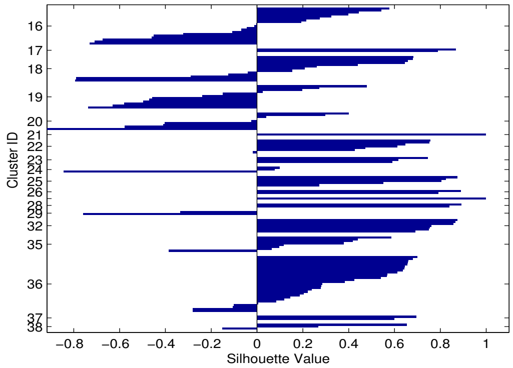


Fig. 12. Silhouette cluster validation index for the 24h feature-set.

remain after the filtering step indicates that the PCA-based method performs sufficiently well (Figure 11).

As part of the cluster validation process, we repeated clustering with a reduced 24-element feature set. Each one of the 24 features corresponded to the mean hourly session arrival rate estimated over all days of the tracing period. This feature set is more compact, while it can still capture the diurnal effect. However, it results in a significant reduction of the *GSu* value, 0.21 vs. 0.39 for the 192-element feature set (12). It comes out that the 192-element feature set reflects better the temporal variation across week days. Averaging daily session arrival rates, the 24-element feature set results in loss of detail in our data with negative impact on clustering.

With the first clustering alternative, combining SVD with the scale factor vector, we obtain 9 clusters. The largest cluster contains 216 buildings, all of which satisfy the filtering criteria for low-utilized buildings. The remaining 8 clusters all appear reasonable; the *GSu* index is 0.76 when all clusters are considered. The second SVD-based clustering alternative produces 26 clusters; eighteen are filtered out when applying our filtering rules corresponding to 181 buildings. The estimated *GSu* index is 0.40. Out of the three approaches, the SVD-scale factor approach results in the best separation between high/low traffic buildings, whereas the one huge cluster produced consists of buildings with low session arrival rates.

B. *Clustering with respect to session-level flow-related variables*

After clustering the buildings according to session arrival rate, we need to model the session-level flow-related variables (see Table II). There are different alternatives for this task in light of the clustering work for session arrivals. One option would be to perform a separate clustering of buildings for each flow-related attribute. However, this approach would give rise to a large total number of clusters and would complicate the modeling effort, effectively canceling the clustering benefits. A variant of this would be to carry out the additional clustering *within* each cluster obtained from the session arrival rate clustering. Another alternative is to consider modeling flow-related variables using the *same* building groups that come out of the clustering on session arrival rates. In our validation analysis, we will compare this last approach against more heuristic groupings (*i.e.*, based on the building-type).

VI. Validation

A. *Methodology*

In this section we evaluate the different modeling alternatives described in Sections III and V. We work with individual buildings and compare how the models' capability to capture the traffic demand at the building level changes as we zoom in/out of the trace data and consider different levels of detail in our modeling.

We consider two alternatives for parameterizing the time-varying Poisson process for the session arrivals: the hourly session arrivals of the specific building we study and the signature of the cluster this building was assigned to. For the flow-related variables, the levels we consider in increasing order of spatial aggregation are the building, cluster, building-type and, for comparison reference purposes, the network level. We combine these alternatives into six scenarios and use them alternately in our simulations to assist the illustration of our main findings and support our discussion. Table V summarizes these scenarios and their requirements in terms of sampling distributions when the whole wireless network is to be modeled. [2]

Given the heavy-tailed session durations, our simulation times are in the order of days rather than hours.

[2] Regarding the numbers in the 3rd column, note that, irrespective of the modeling scenario, we always need to model four variables: the session arrival process and the three flow-related variables.

TABLE V

MODELING ALTERNATIVES-SCENARIOS FOR SIMULATION VALIDATION.

Modeling scenario	Description	Sampling distributions
bldg-bldg	Both session arrivals and flow-related variables are modeled after bldg-specific data for the whole trace duration	4*N N : number of bldgs
bldg-bldg (day)	The same with bldg-bldg, only now different distributions are derived for each day of the monitoring period	4*N*D D : number of days
bldg-bldgtype	Session arrivals are modeled after bldg-specific data and flow-related variables over data aggregated at bldg type level	N + 3*M, M : num. of bldg types
cluster-bldg	The cluster signature is used for session arrival modeling, whereas the distributions for flow-related variables are drawn from building-specific data	C + 3*N, C : number of clusters
cluster-cluster	Both session arrivals and flow-related variables are modeled after clustering	4*C
bldg-network	Bldg-specific data for session arrivals, network-wide distributions for the flow-related variables	N+3

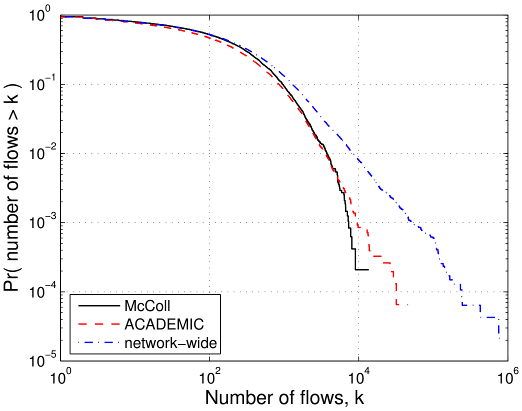


Fig. 13. Number of flows per session : ccdf under different building-grouping alternatives.

We have implemented the thinning process described in [19] to simulate the time-varying Poisson process for the session arrivals.

We compare synthetic traffic against traces with respect to building-level 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 building under study. We examine first-order and second-order statistics of the flow interarrival process and hourly flow arrival counts. We present results for two buildings, one academic (McColl) and one residential (Hinton James). They are two of the busiest campus buildings and represent the two main building types. In the same time they exemplify the value of heuristics (McColl) and clustering techniques (Hinton James) for achieving a good trade-off between model accuracy and scalability.

B. *McColl academic building*

McColl is the busiest campus building in terms of session arrivals. Irrespective of the clustering approach followed in Section V, the McColl building does not group with other buildings but rather forms a separate cluster on its own. Therefore, the cluster signature coincides with the building hourly session arrival rate and the modeling alternatives are only relevant to the set of distributions for the flow-related variables, which are listed in VI

A first view of the "noise" that averaging introduces is given in Figure 13. The plot compares the cumulative empirical distribution function of the in-session number of flows for the McColl building with those estimated for all academic buildings and network-wide. The deviation between the curves increases with the degree of spatial aggregation of data. The way these discrepancies affect the aggregate flow-related metrics we described in Section VI-A is summarized in Figures 15-17

With respect to aggregate flow interarrivals, the synthetic traffic generator tracks most closely the trace when we model the in-session flow number and flow interarrivals separately for each one of the three days of simulation time. Considering a single set of distributions over the whole trace, does not give better results than when using the aggregate distributions for academic buildings. This implies that the averaging in the time-dimension may cancel out the benefit of getting higher spatial resolution out of the trace data. In the same time, staying at the building-type level gives us comparable precision with that obtained when zooming into the building-specific data. Figure 15 clearly suggests that reuse of the network-wide distributions for modeling traffic demand at finer spatial scales is not an attractive alternative. Looking at the autocorrelation process in Figure 16, one notes that the bldg-bldgtype curve is not much worse than the one corresponding to the bldg-bldg scenario. In fact, the former seems to underestimate less the short-term autocorrelation than the latter.

Finally, inferior in absolute terms is the match for the hourly flow counts, which is the most demanding metric. Figure 17 plots the averages over 40 simulation runs along with their 95% confidence intervals. In this case, further improvement would be obtained by modeling the flow-related variables over shorter time periods than over the full monitoring period or a day. In fact, the standard practice is to focus the modeling attention on short time windows where the building activity experiences its peak (busy hour). In any case, the aggregation along the building type performs only marginally worse than

TABLE VI

MCCOLL ACADEMIC BUILDING : PARAMETERS FOR THE FLOW-RELATED VARIABLE DISTRIBUTIONS UNDER VARIOUS LEVELS OF SPATIAL AGGREGATION IN MODELING

	McColl	Academic	Network-wide
Flow number/session (BiPareto)	$\alpha = 0.09, \beta = 2.69,$ $c = 1026.37, k = 1$	$\alpha = 0.11, \beta = 2.17,$ $c = 713.85, k = 1$	$\alpha = 0.09, \beta = 1.49,$ $c = 585.4, k = 1$
Flow interarrivals/session (Lognormal)	$\mu = -1.69, \sigma = 3.01$	$\mu = -1.65, \sigma = 2.99$	$\mu = -1.49, \sigma = 2.92$

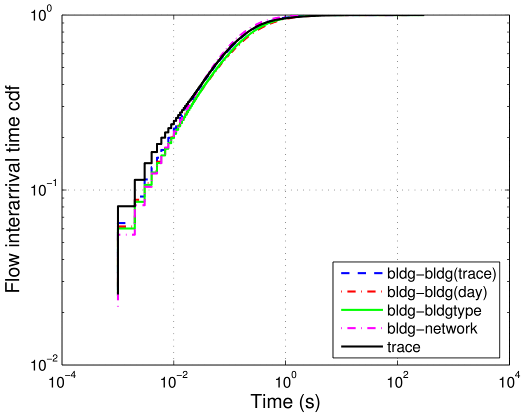


Fig. 14. McColl building: aggregate flow interarrivals cdf

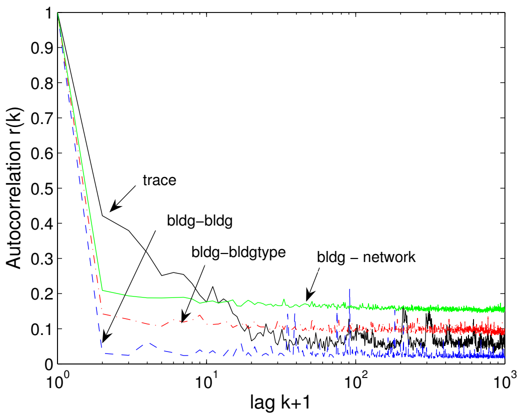


Fig. 16. McColl building: autocorrelation of aggregate flow interarrivals

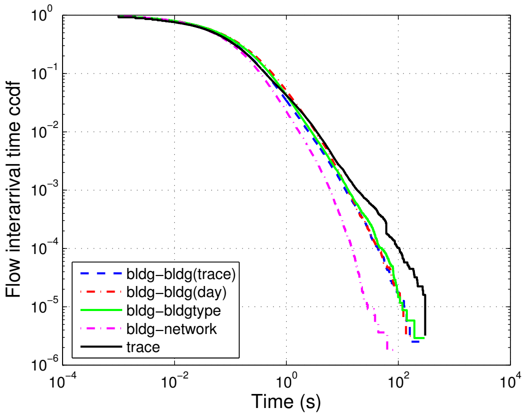


Fig. 15. McColl building: cumulative empirical distribution function of aggregate flow interarrivals

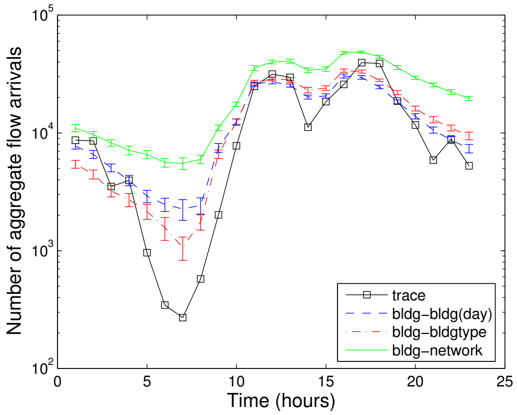


Fig. 17. McColl building: aggregate hourly flow arrivals

the bldg-bldg scenario. The required number of sampling distributions for modeling each campus building under the bldg-bldg scenario would be 4*N*D = 3000, for N = 250 and D = 3. When all buildings of the same type are modeled after a common set of distributions for flow-related variables, their number is reduced down to N + 3*M = 274, for M = 8.

C. *Hinton James residential building*

Contrary to McColl, the Hinton James building was clustered together with other buildings under all clustering alternatives described in Section V. We consider for further analysis the cluster that came out of the SVD-scale factor technique, which is the one that gave the highest global silhouette index amongst the three alternatives. This cluster comprises three other buildings, two of the social and one of the library type. Therefore, one additional modeling alternative now is to consider the cluster signature instead of its hourly session arrival rate for modeling the session arrivals (Figure 18).

We consider four alternative scenarios in our simulator. In two of them we model session arrivals after the building hourly session arrivals and in the other two we use the cluster signature. For the flow-related variables, we have three alternatives for getting the respective sampling distributions: consider only the building-specific data, aggregate over data from all four buildings in

TABLE VII

HINTON JAMES RESIDENTIAL BUILDING : PARAMETERS FOR THE FLOW-RELATED VARIABLE DISTRIBUTIONS UNDER VARIOUS LEVELS OF SPATIAL AGGREGATION IN MODELING

	Hinton James	Residential	Clustering	Network-wide
Flow number/session (BiPareto)	$\alpha = 0.08, \beta = 1.95,$ $c = 1357.34, k = 1$	$\alpha = 0.08, \beta = 1.34,$ $c = 968.37, k = 1$	$\alpha = 0.081, \beta = 1.87,$ $c = 680.18, k = 1$	$\alpha = 0.09, \beta = 1.49,$ $c = 585.4, k = 1$
Flow interarrivals/session (Lognormal)	$\mu = -1.62, \sigma = 2.99$	$\mu = -1.44, \sigma = 2.8$	$\mu = -1.62, \sigma = 2.97$	$\mu = -1.49, \sigma = 2.92$

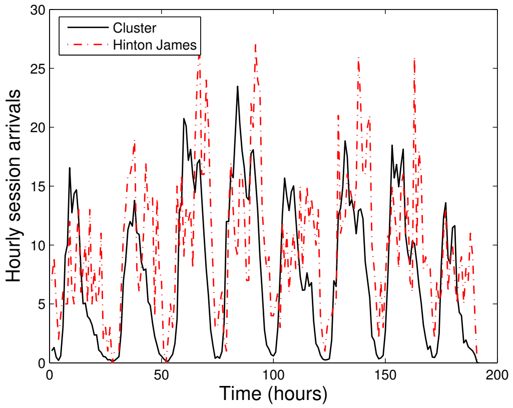


Fig. 18. Hourly session arrivals : Hinton James vs. cluster signature

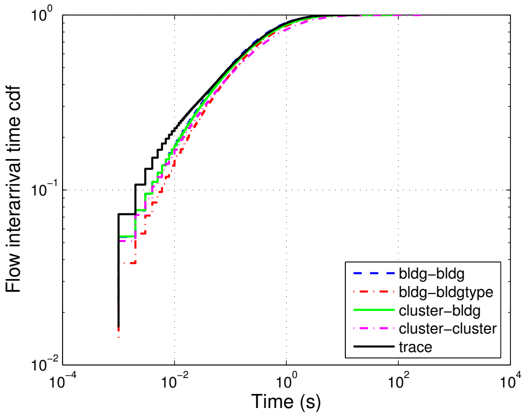


Fig. 19. Hinton James building: empirical distribution function of aggregate flow interarrivals

the cluster and, aggregate over data from all residential buildings in campus. The parameters of the sampling distributions in each case are given in Table VII.

Figures 19-22 summarize the relative performance of the four alternatives. Now, the precision vs. aggregation level trade-off is even more clear than with the McColl building. Interestingly, the cluster-bldg combination presents a good compromise with matching score too close to the one obtained when we consider the specific building session arrival rates.

The bldg-bldgtype alternative is inferior to the cluster-bldg one, implying that modeling in-session flow number and flow interarrivals by simply aggregating data from all residential buildings has some cost. However, it is still better than reusing the clustering results obtained for the hourly session arrival rates to the modeling of the flow-related variables. This becomes clear in Figures 19 and 20, which show that when this happens the mismatch is the worst of all scenarios. It comes out from these plots that the building type can be most useful in grouping buildings with respect to the session-level flow-related variables we model.

The relative performance of the four scenarios is preserved with respect to the second-order statistics (Figure 22) and the flow-counts (Figure 21). In the second case, as with the McColl building, the implication is that additional detail in the time domain may be necessary to get higher precision. Nevertheless, considering the cluster signature instead of the per-building session arrival rates gives almost identical performance. Again, this happens at the benefit of scalability, since the required number of sampling distributions (parameter sets) to be input to the simulator is C+3*N = 770, for C = 20 versus 4*N = 1000, respectively. Even better scalability in this case would be achieved under a combination cluster-bldgtype, *i.e.*, if we used the clustering results to model the session arrival process and aggregate data at the building-type level for the flow-related variables. The required number of sampling distributions would be C+3*M = 44, resulting in an impressive reduction of complexity in the simulator.

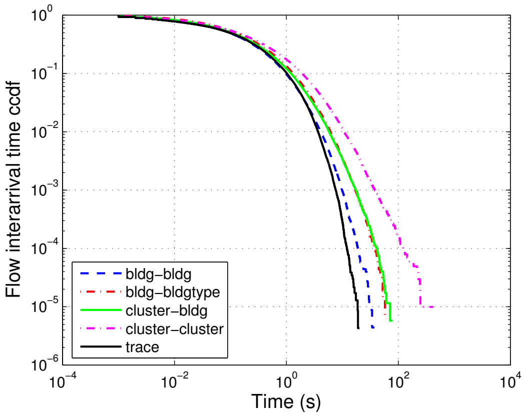


Fig. 20. Hinton James building: cumulative empirical distribution function of aggregate flow interarrivals

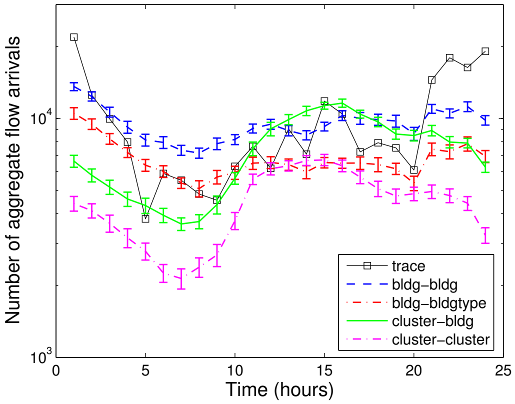


Fig. 21. Hinton James building: aggregate hourly flow arrivals

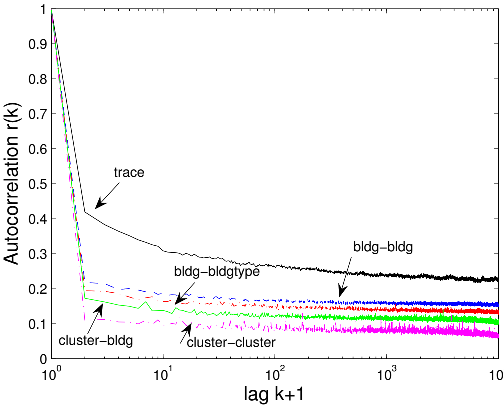


Fig. 22. Hinton James building: autocorrelation of aggregate flow interarrivals

VII. Related work

Measurement-based studies of WLAN traffic demand have a much shorter history than those for wired networks (for example, [2]–[4], [13], [20]). High-level observations about the temporal and spatial variation of the traffic demand appear in a number of papers [5]–[7], [21], [22], which have drawn measurement data from different types of wireless infrastructures: campus WLANs [5], [21], [22], enterprise WLANs [6], conference hotspots [7]). Traffic load diurnal/weekly periodicities have been noted in [5], [6], [21], [23]. Almost all studies describe traffic demand variations amongst the monitored APs, with [5], [6], [14] also describing building and building-type dependencies.

Nevertheless, the first study that addressed the WLAN traffic modeling at higher detail is the one by Meng *et al.* [8]. They use measurement data from the University of Dartmouth campus WLAN 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 flow sizes 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.

In earlier work in [9], we look into traffic demand at network-level. Contrary to [8], the non-stationarity of traffic workload is captured at the session- rather than 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* variation of the workload. In this study, we draw on the work in [9], only now we take a closer look at the spatial variation of traffic demand. We work with buildings rather than APs and propose ways that enable scalable yet accurate modeling of the traffic demand in the network.

VIII. Conclusions

Our paper addresses the problem of traffic demand modeling in large wireless networks. We emphasize on the spatial dimension of the traffic load variation looking at various scales of spatial aggregation in the wireless network.

In earlier work [9] we proposed a hierarchical modeling framework for aggregate network-wide traffic demand drawing on wireless sessions and network flows. In this paper, we derive two notable results related to it. Firstly, we find out that the statistical distributions proposed for network-wide traffic demand, *i.e.*, time varying Poisson process for session arrivals, BiPareto for in-session flow numbers and flow sizes, and Lognormal for in-session flow interarrivals, are valid over two different monitoring periods, spaced one year apart. Secondly, the same distributions apply when we look at traffic at finer spatial scales, such as the single building or groups of buildings with similar usage. Given the second result, we promote buildings as the primary entities for traffic demand modeling in the spatial dimension. Working at building level circumvents several problems emerging when working at AP-level: non amenability to statistical processing, higher sensitivity of monitored traffic variables to the short-term propagation conditions, lack of scalability.

We elaborate on this last aspect and propose a novel methodology for coming up with models that scale with the size of the network whilst preserving modeling accuracy. The methodology involves the use of heuristics and statistical clustering techniques to group buildings for traffic demand modeling purposes. By way of example, we show that both of them can benefit the traffic modeling task. Interestingly, the two approaches are complementary. Heuristical segregation of buildings scores better with flow-related variables, whereas clustering combined with PCA and SVD performs better with the modeling of session arrivals, where the requirement is to group buildings with similar volume and not only pattern of arrivals.

To encourage further experimentation along the lines drawn in this paper, we have made our datasets and tools available to the research community [10].

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".

References

[1] V. Paxson and S. Floyd, "Wide area traffic: the failure of poisson modeling," *IEEE/ACM ToN*, vol. 3, no. 3, pp. 226–244, 1995.

[2] 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.

[3] 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*, Santa Clara, CA, United States, June 2000, pp. 166–177.

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

[5] 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.

[6] 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.

[7] 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.

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

[9] F. Hernandez-Campos, M. Karaliopoulos, M. Papadopouli, and H. Shen, "Spatio-temporal modeling of traffic workload in a campus WLAN," in *Second Annual International Wireless Internet Conference*, Boston, MA, USA, 2006.

[10] UNC-FORTH Archive of Wireless Traces, Models, and Tools. [Online]. Available: http://www.cs.unc.edu/Research/mobile/datatraces.htm

[11] "Cisco Aironet AP specifications," Product information sheet. [Online]. Available: http://www.cisco.com/en/US/products/hw/wireless

[12] L. Yang, P. Zerfos, and E.Sadot, "Architecture taxonomy for control and provisioning of Wireless Access Points (capwap)," RFC 4118, June 2005.

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

[14] 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.

[15] I. T. Jolliffe, *Principal Components Analysis*. New York: Springer Verlag, 2002.

[16] H. C. Rosemburg, *Cluster Analysis for Researchers*. Belmont, CA, United States: Lifetime Learning Publications, 1984.

[17] R. Mojena, "Hierarchical grouping methods and stopping rules: An evaluation." *Computer Journal*, vol. 20, p. 359363, 1977.

[18] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

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

[20] V. Paxson, "Empirically-derived analytic models of wide-area TCP connections," *IEEE/ACM ToN*, vol. 2, no. 4, pp. 316–336, August 1994.

[21] D. Tang and M. Baker, "Analysis of a local-area wireless network," in *Proc. of ACM MobiCom*, Boston, MA, United States, Aug 2000, pp. 1–10.

[22] 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.

[23] 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.