Session Lengths and IP Address Usage of Smartphones in a University Campus WiFi Network: Characterization and Analytical Models

Xian Chen, Lester Lipsky, Kyoungwon Suh[†], Bing Wang, Wei Wei
University of Connecticut, [†]Illinois State University

***Abstract*—Smart mobile handheld devices (MHDs) are being adopted at a fast speed. Compared to wireless non-handheld devices (NHDs), MHDs tend to be more mobile and can be used more opportunistically. In this paper, we study two important network usage characteristics of MHDs, namely session lengths and IP address usage, in a university campus WiFi network. Specifically, we analyze two five-week long DHCP traces collected from the network, characterize session lengths of MHDs, and develop two hyper-exponential models to capture the distribution of session lengths. We further characterize the IP address usage of MHDs, and develop two analytical models to predict the number of concurrent IP addresses that are being used by MHDs at one point of time. Goodness of fit tests indicate that our analytical models of session lengths provide good fit, and evaluation results demonstrate that the predictions from our models for IP address usage are accurate. Our results provide important insights on managing MHDs as they are being adopted rapidly in WiFi networks.**

I. Introduction

Smartphones are one of the most rapidly adopted consumer technologies of all time [3]. A recent PEW study indicates that smartphone owners have outnumbered basic phone users in 2012 [2]. Market analysts predict that within five to seven years, smartphones will be the only mobile phones used in the United States [3], [4]. In campus scenarios where WiFi networks are densely deployed and well provisioned, people often prefer to connect their smartphones to WiFi networks (instead of cellular networks) for Internet access, because of their higher bandwidth, lower delay, lower cost, and lower energy consumption [8], [14], [12], [7], [31], [23]. Compared to laptops and desktop PCs, network usage characteristics of smartphones may differ significantly because smartphones are more mobile and are being used more opportunistically — the small physical forms of smartphones allow them to be carried by their owners most of the time, and used whenever possible. As smartphones are being adopted rapidly, it is important to understand their network usage characteristics, which can have immense implications on network management.

In this paper, we study network usage characteristics of smartphones in a university campus WiFi network. We consider both smartphones and tablets, and broadly referred them as smart mobile handheld devices (MHDs). Specifically, we focus on two network usage characteristics of MHDs, *session length* and *IP address usage*, that are important for network management. Session length represents how long a client is connected to the network. Characterizing session length is important for studying mobility, planning wireless infrastructure deployment, designing protocols for wireless applications and services, and analyzing their performance. In addition, session length characteristics can be used by access points, proxies, and servers to prepare handoffs, share clients or traffic load with each other, and ensure a better service quality [28]. Understanding IP address usage is important for network planning. The large number of MHDs has increased the demand on IP addresses, imposing significant stress on IP address management. Accurate prediction of IP address usage of by the MHDs can help network administrators to take proactive actions to satisfy future IP address demands.

While the literature on MHDs and non-handheld devices (NHDs) in WiFi networks is extensive (see Section VI), there is little study on session lengths and IP address usage of MHDs in campus WiFi networks. Our study analyzes two five-week long DHCP (Dynamic Host Configuration Protocol) traces collected from the University of Connecticut (UConn) campus WiFi network. We focus on the results on MHDs, and when necessary, present the results on NHDs for comparison. Our main contributions are as follows.

- We characterize session lengths of MHDs, and develop two hyper-exponential models to capture the distribution of session lengths. Goodness of fit tests indicate that both models provide good fit. In addition, we find hyper-exponential models provide better match to the empirical data than BiPareto models [21], [28], [27] that have been proposed for NHDs.
- We characterize the IP address usage of MHDs, and develop two analytical models to predict the number of concurrent IP addresses that are being used by MHDs at one point of time. Evaluation using the data collected from UConn campus WiFi network demonstrates that predictions from our models are accurate. To the best of our knowledge, we are the first to propose analytical models for IP address space usage when the IP addresses are managed by DHCP.

The rest of the paper is organized as follows. Section II presents background on DHCP. Section III describes our methodology for collecting and analyzing data. Sections IV and V present the results on characterizing and modeling

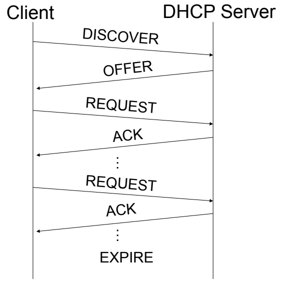


Fig. 1. An example DHCP message exchange.

session lengths and IP address usage of MHDs, respectively. Section VI briefly reviews related work. Finally, Section VII concludes the paper and presents future work.

II. Background on DHCP

DHCP is the *de facto* protocol for managing IP addresses in campus and enterprise networks. When connecting to a WiFi network, a wireless device (MHD or NHD) uses DHCP to acquire IP addresses. We next briefly describe DHCP; more details can be found in [13].

In general, DHCP allocates an IP address from a predefined IP address pool to a host that joins the network, and reclaims that IP address when the lease time expires. Fig. 1 illustrates an example DHCP message exchange between a host and a DHCP server. When a host connects to a network, it first broadcasts a *discover* message, and each of the available DHCP servers replies with an *offer* message that contains an offered IP address[1]. Among all of the offered IP addresses, the host chooses one and sends a unicast *request* message to the DHCP server. After that, the server confirms the IP allocation via an *ACK* message with an IP lease time. The host needs to renew the lease by sending a *request* message again after half of the IP lease time [13], [25]. If the host does not send a *request* message before the expiration time, the server reclaims that IP address back to the pool, and logs an *expire* message. When leaving the network, the host can send a *release* message to the server (optional). An important parameter for DHCP configuration is the default IP lease time, which specifies by default, how long a host can lease an IP address. In UConn, the default IP lease time for the WiFi network is 30 minutes.

III. Data Collection and Methodology

A. Data Collection

Our data are collected from the University of Connecticut (UConn) campus network. The campus has two DHCP servers, both running a modified version of ISC-DHCP, and logging all DHCP related messages into a central database. Since UConn uses separate IP address pools for wired and wireless devices, we can extract the DHCP trace from the database that only contains messages for wireless devices and use it for our study.

Using the above methodology, we obtain two sets of DHCP records of wireless hosts, one from February 20 to March 25, 2012, and the other from August 20 to September 23, 2012. Each dataset contains five weeks of records. The first dataset is 26GB, containing 209M records. The second dataset is 34GB, containing 279M records. We henceforth refer these two datasets as Spring and Fall dataset, respectively. Each entry in the datasets corresponds to a DHCP message with an MAC address, the time of the message, and other information (depending on the type of the message). We use host MAC address to uniquely identify a host.

In the Spring dataset, the week between March 12 and March 18 is a spring break week; in the Fall dataset, the week between August 20 and August 26 is the last week of summer. For these two weeks, the size of the trace is significantly smaller (around 90% smaller) than that of a regular week. In the rest of the paper, we focus on the eight regular weeks, excluding these two weeks that have significantly less records.

B. Methodology

The DHCP trace contains a mixture of records for MHDs (e.g., iPhones, iPod touches, Android phones, Windows phones, and Blackberry phones) and NHDs (e.g., Windows laptops and MacBooks). Since MHDs and NHDs may have different network usage characteristics (as we shall see), we first describe a methodology that determines whether a host is an MHD or NHD. We then describe a methodology to obtain *session length*, i.e., how long a client is connected to the network, and a methodology to obtain *IP address allocation duration*, i.e., how long an IP address is being allocated to a host. In the rest of the paper, we use user, host, and client interchangeably.

1) Determining Host Type: Two approaches have been used to differentiate MHDs and NHDs: one approach uses the fingerprints (e.g., host name, vendor name, parameter in request, and Organization Unique Identifier) in DHCP packets [30]; the other approach uses keywords in the user-agent field in HTTP headers [24], [18], [10]. We use a combination of the above two approaches. This is because the first approach can be applied directly to DHCP records (which we have collected), while the second approach requires collecting additional data (i.e., HTTP headers). On the other hand, our DHCP records only contain host names, one feature of the set of features used in [30]. We therefore use a simplified version of the first approach (i.e., only using host names) to classify hosts, and use the second approach to verify the accuracy of classification results thus obtained since user-agent fields in HTTP headers are more reliable than host names (the former are determined by operating systems while the latter are determined by users).

Specifically, we first use keywords in host names of DHCP packets to differentiate MHDs and NHDs. Keywords for MHDs include iPhone, iTouch, iPad, Android, Blackberry and so on; keywords for NHDs include PC, Windows, Macbook and so on. Using this simple approach, we can determine host types for over 90% of the hosts. We then collect HTTP traffic, use the latter approach to differentiate MHDs and NHDs, and compare the results with those obtained from the approach based on host names. Specifically, we capture one week, from

[1] Most DHCP relay agents can forward DHCP configuration requests, eliminating the requirement of setting up a DHCP server on every subnet [1].

March 19 to March 24, 2012, of HTTP traffic (up to 500 bytes for each packet so that HTTP header is captured), and use the user-agent field in HTTP headers to identify the OS type of each IP address [24], [18], [10]. We then correlate the HTTP headers and the corresponding DHCP traces using the source IP address and the time so as to obtain a mapping between an IP address and an MAC address. In this way, we can obtain the OS type for each MAC address, which can be used to easily determine whether a host is an MHD or NHD. The results using this approach are consistent with the results by using host names (there is only 0.39% discrepancy).

In the rest of the paper, host classification is through the approach that is based on host names[2]. From the two DHCP traces (Spring and Fall dataset), we identify about 48,000 unique NHDs and 18,000 unique MHDs.

2) Estimating Session Length: Consider a host. Let t_s denote the time when the host starts to own an IP address, and t_e denote the time that the host does not use this address any more. Then the session length is $t_e - t_s$. From Fig. 1, we see t_s is the time of the first ACK message after the *discover* message, which can be easily determined from the DHCP trace. Determining t_e is much more challenging. This is because *release* and *expire* messages are optional, and indeed we find that they are not frequently used in practice. In addition, we find that a host may join the network again before the current IP lease expires. Summarizing the scenarios we observe from the DHCP trace, we propose the following rules to determine t_e.

R1: If the IP address is assigned to another host at time t, which is earlier than the lease expiration time, we set $t_e = t$.

R2: If the host sends a *discover* message at time t, which is earlier than the lease expiration time, we set $t_e = t$.

R3: If an *expire* or *release* message is observed for the host, then t_e is the time when the message is observed.

R4: If no *expire* or *release* message is observed before the expiration time, then we use the lease expiration time as t_e[3].

R1 corresponds to scenarios where the DHCP server assigns the IP address to another host even before the current lease expires, which should happen very rarely in a normally running network (See Table I). R2 corresponds to scenarios where the host initiates another IP lease period before the current lease expires, and hence the current IP lease period should be terminated. R3 corresponds to scenarios where t_e is indicated by an explicit message, while R4 corresponds to scenarios where t_e is determined based on IP lease expiration time in the absence of any explicit message. The above rules are more comprehensive than those in [25] (which only consider the last two cases).

Table I presents the percentage of scenarios following each of the four rules described above. The results are for the Fall dataset (The results for the Spring dataset are similar, and are omitted). First, we observe for MHDs, the percentage of scenarios following R2 is significantly larger than that for NHDs, indicating that MHDs connect to the network much more frequently. Second, R3 and R4 in total only account for 34.4% and 77.6% of the scenarios for MHDs and NHDs, respectively. Hence it is important to consider rules R1 and R2 in UConn network (only applying R3 and R4 as in [25] is insufficient for UConn network). Last, the large percentage of the scenarios following R4 for NHDs indicates that many clients do not send *expire* or *release* messages to the DHCP server, an observation consistent with that in other studies [25]. On the other hand, the percentage of the scenarios following R4 for MHDs is much lower (only 1.3% versus 14.2% as in NHDs).

TABLE I
ESTIMATING SESSION LENGTH: THE PERCENTAGE OF SCENARIOS FOLLOWING EACH OF THE FOUR RULES (FALL DATASET).

	R1	R2	R3	R4
MHD	0.02%	66.1%	33.1%	1.3%
NHD	0.03%	22.4%	63.4%	14.2%

The above method for determining session lengths may provide an overestimate in certain scenarios (e.g., when a host has departed a network while the IP address has not been reclaimed due to lack of *release* message). Combining networking traffic (which shows when a host does not have any network traffic for an extended period of time) with DHCP records can provide more accurate estimates of session lengths [30], particularly for short session lengths. On the other hand, we believe the overestimate from our approach does not affect our conclusion that session lengths for MHDs tend to be much shorter than those of NHDs, and is better matched using hyper-exponential models instead of BiPareto models (see Section IV). Obtaining more accurate estimates of session lengths is left as future work.

3) Estimating IP Address Allocation Duration: Consider an IP address. Let t denote the time when the IP is assigned to one host, and t' denote the time that the DHCP server reclaims the address from that host. Then $t' - t$ represents the allocation duration of the IP address. We determine t as the time of the first ACK message after the *discover* message, and use rules R1, R3, and R4 to determine t' (R2 does not apply since a new *discover* message from a host does not mean that the IP address allocated to the host has been reclaimed, see the example below).

Note that for a host that obtains an IP address, its session length and IP address allocation duration can be different. In particular, IP address allocation duration may be longer than session length since a DHCP server does not reclaim the IP from a client until the IP lease time expires or the client sends

[2] Note that although the simple approach works for our campus, it may not work well for other campuses given that host names are not reliable. In fact, even user-agent fields in HTTP headers can be spoofed. Developing statistical techniques that rely on intrinsic features, and hence cannot be easily spoofed, to reliably differentiate MHDs and NHDs is an interesting future direction that we will pursue further.

[3] Since we focus on when the IP will be reclaimed, we use the expiration time instead of the last ACK message time plus half of the default renew period [25].

out a *release* message explicitly. For instance, assume that a client obtains an IP address, a, at time t_1, and the expiration time of the IP address is t_2. Suppose the client moves to another location and obtains a different IP at time t_3 ($t_3 < t_2$), and subsequently leaves the network at time t_4. Then for the client, the session length of using IP address a is $t_3 - t_1$, while the allocation duration of IP address a is $t_2 - t_1$. From the DHCP server's perspective, that client owns two IP addresses simultaneously from t_3 to t_2.

IV. Characterizing and Modeling Session Lengths

We first present session length distribution of MHDs obtained from our dataset, and then present two analytical models for session lengths, followed by validation of the models.

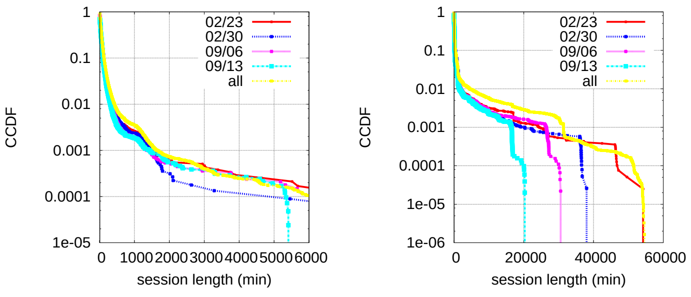
A. Session Length Distribution

For both MHDs and NHDs, we observe peak arrival rates (i.e., the rate hosts connect to the WiFi network) between 9am and 3pm on each day (see Section V-A). We next characterize the lengths of the sessions during that time period. Specifically, for users arriving between each of the two hour intervals between 9am and 3pm (i.e., [9-11]am, [11am-1pm], [1-3]pm), we obtain their corresponding session length distribution. The results (not plotted) show that the distributions are similar, indicating that session lengths are not time sensitive. Therefore, for each day, we only obtain the session length distribution for all hosts arriving between 9am and 3pm. We also obtain the overall session length distribution, i.e., the distribution obtained from all the sessions in the two datasets.

Fig. 2(a) plots the CCDF (Complementary Cumulative Distribution Function) of session lengths of MHDs on four Thursdays, two from the Spring dataset and two from the Fall dataset, and the overall session length distribution. We observe that all the session length distributions are similar: the distribution of the session lengths collected in a day is similar for the Spring and Fall semester, and is similar to the overall distribution.

For comparison, Fig. 2(b) plots the session length distribution of NHDs. Observe that the tail of the distribution for NHDs can be much longer than that of MHDs. In addition, as expected, session lengths of MHDs tend to be shorter than those of NHDs. For instance, over all the data, the median of session length for MHDs is 32.8 minutes, while the median session length for NHDs is 64.8 minutes. In addition, 8.3% of the sessions for NHDs while only 0.7% of the sessions for MHDs stay longer than 500 minutes.

Recall that 30 minutes is the default IP lease time for UConn network. Therefore, users that stay in the network shorter than 15 minutes will own their IP addresses for 30 minutes, leading to a large percentage of session length as 30 minutes (9.4% and 10.9% for MHDs and NHDs, respectively). This causes a jump in the CCDF curve (not clearly visible in Fig. 2).



(a) MHDs

(b) NHDs

Fig. 2. Session length distribution for four Thursdays and for all the data (including both the Spring and Fall datasets).

B. Analytical Models

Visually inspecting the CCDF of session lengths of MHDs, we observe that it can be fit by two or three straight lines, corresponding to different portions of the curve. Therefore, we use two-stage or three-stage hyper-exponential distribution to model session length. We also tried BiPareto distributions as proposed [21], [28], [27] and found the goodness of fit is worse than those from the two hyper-exponential models, which might be due to the much shorter tail of the session length distribution of MHDs compared to that of NHDs. In the following, we only present the hyper-exponential models.

Let $f(x)$ denote the probability density function for session length of MHDs. Incorporating the significant fraction of session lengths that are of the default IP lease time, we represent $f(x)$ as

$$f(x) = (1-b)g(x) + b\delta(x - x_v) \tag{1}$$

where x_v is the default IP lease time, $b \in [0,1]$ represents the fraction of session length of x_v, $\delta(\cdot)$ is the indicator function, and $g(x)$ is a two-stage or three-stage hyper-exponential distribution.

1) Two-Stage Hyper-exponential Model: In this model, $g(x)$ is represented as

$$g(x) = \mu_1 p e^{-\mu_1 x} + \mu_2(1-p)e^{-\mu_2 x} \tag{2}$$

where $\mu_1 > 0, \mu_2 > 0$, and $p \in [0,1]$ are parameters in the model. We next describe how to obtain these parameters.

Let random variable X denote session length. Then the first two moments of X are

$$E(X) = (1-b)\alpha + bx_v, \tag{3}$$

$$E(X^2) = (1-b)\beta + bx_v^2, \tag{4}$$

where

$$\alpha = \frac{p}{\mu_1} + \frac{1-p}{\mu_2}, \tag{5}$$

$$\beta = \frac{2p}{\mu_1^2} + \frac{2(1-p)}{\mu_2^2}. \tag{6}$$

We can obtain $E(X), E(X^2)$ and b directly from the trace, which can then be used to solve for α and β using equations (3) and (4). After that, we have two equations (5) and (6) with three unknowns, μ_1, μ_2, and p, which cannot be solved to obtain μ_1, μ_2, and p directly. On the other hand, many techniques can be used to create a hyper-exponential distribution that matches the given first two moments. We use the following procedure to obtain μ_1, μ_2, and p [5].

- Calculate $CV^2 = \frac{\beta - \alpha^2}{\alpha^2}$.
- Calculate $p = \frac{1}{2}(1 - \sqrt{\frac{CV^2-1}{CV^2+1}})$. This requires $CV \geq 1$, which is true in our scenario.
- Set $\mu_1 = \frac{2p}{\alpha}$, and $\mu_2 = \frac{2(1-p)}{\alpha}$.

2) Three-Stage Hyper-exponential Model: In this model, $g(x)$ is represented as

$$g(x) = \mu_1 p_1 e^{-\mu_1 x} + \mu_2 p_2 e^{-\mu_2 x} + \mu_3 p_3 e^{-\mu_3 x} \qquad (7)$$

where $\mu_1 > 0, \mu_2 > 0, \mu_3 > 0$, and $p_1, p_2, p_3 \in [0, 1]$ are the parameters of the model, and $p_1 + p_2 + p_3 = 1$. To obtain the parameters in (7), we use the iterative Feldmann and Whitt procedure [17] as follows. Given a desired number of phases, it iteratively computes the parameters of each so as to match certain points on the CDF. The procedure starts from the tail and works its way towards the origin, taking into account the phases that have already been defined, and matching what is left over.

Let a set $0 < c_3 < c_2 < c_1$ of points divide the range of interest into exponentially-related subranges. Specifically, c_1 represents the highest values that are of interest, and c_3 represents the smallest values that are of interest. The ratio c_i/c_{i+1} is set to a constant c. We select $c = \sqrt{c_1/c_3}$. Let $q = \sqrt{c}$, where qc_1 must not be larger than the highest data point. Then, $\bar{F}(x) = \Pr(X > x)$ can be calculated as follows.

- Initially we match the first phase ($i = 1$) to the tail of the given data. In other words, we have $\bar{F}_1(x) = \bar{F}(x)$.
- In general, in step i we match the ith phase to the tail of the remaining $\bar{F}_i(x)$. An exponential phase has two parameters, p_i and μ_i. To find the values of these two parameters, we match $\bar{F}(x)_i$ at two points: c_i and qc_i. For each phase i, $i = 1, 2$,

$$\begin{aligned} p_i &= \bar{F}_i(c_i) e^{\mu_i c_i} \\ \mu_i &= \frac{1}{(1-q)c_i} \ln \frac{\bar{F}_i(qc_i)}{\bar{F}_i(c_i)} \end{aligned} \qquad (8)$$

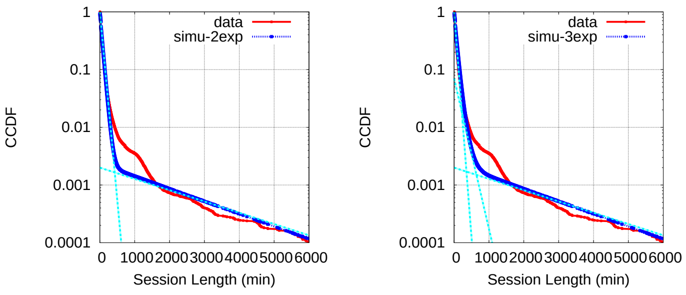
  where

$$\bar{F}_i(c_{i+1}) = \bar{F}(c_{i+1}) - \sum_{j=1}^{i} p_j e^{-\mu_j c_{i+1}}$$

$$\bar{F}_i(qc_{i+1}) = \bar{F}(qc_{i+1}) - \sum_{j=1}^{i} p_j e^{-\mu_j qc_{i+1}}$$

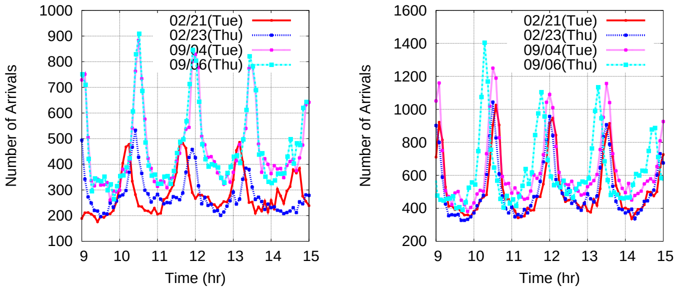
- For the last phase ($i = 3$), the procedure is different. This is to satisfy that $p_1 + p_2 + p_3 = 1$.

$$\begin{aligned} p_3 &= 1 - p_1 - p_2 \\ \mu_3 &= \frac{-1}{c_3} \ln \frac{\bar{F}_3(c_3)}{p_3} \end{aligned}$$



(a) Two-stage hyper-exp model (b) Three-stage hyper-exp model

Fig. 3. Two-stage and three-stage hyper-exponential models for session lengths of MHDs. For ease of illustration, we also plot the straight lines corresponding to the models.



(a) MHDs (b) NHDs

Fig. 4. The number of arrivals on February 21 (Tuesday), February 23 (Thursday), September 4 (Tuesday), and September 6 (Thursday).

C. Model Validation

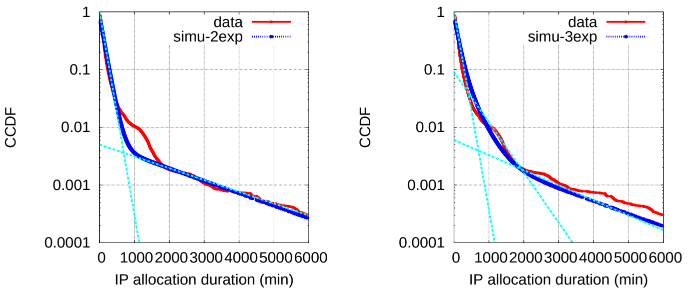
We next validate the models for session length distribution. Figures 3(a) and (b) plot respectively the results of the two-stage and three-stage hyper-exponential models along with the empirical distribution obtained from our data (0.34M samples). In general, we observe a good match from both two-stage and three-stage Hyper-exponential models. Furthermore, we use Kolmogorov-Smirnov test to evaluate the goodness of fit (we do not use Pearson's chi-squared test since it is very sensitive to the binning [21]). The goodness of fit for the two-stage and three-stage models is 0.08 and 0.06, respectively, indicating that both models provide a good fit, and the three-stage model slightly outperforms the two-stage model.

V. Characterizing and Modeling IP Address Usage

In this section, we first present measurement results on two factors that affect IP address usage of MHDs, namely user arrival patterns and IP address allocation duration. We then characterize and model IP address usage of MHDs, validate the models, and use a case study to demonstrate that the models are helpful for network administrators to predict IP address usage in the future.

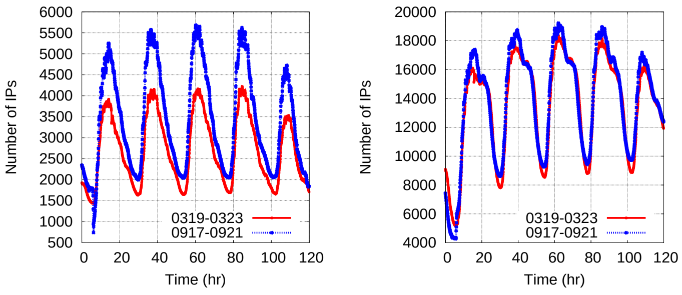
A. User Arrival Patterns

To measure user arrivals, we divide time into five-minute intervals, and obtain the number of user arrivals based on the number of *discover* messages in each interval (since the first



(a) Two-stage hyper-exp model (b) Three-stage hyper-exp model

Fig. 5. For the IP addresses allocated to MHDs, the empirical distribution and the results using the two-stage and three-stage hyper-exponential models for IP address allocation duration. For ease of illustration, we also plot the straight lines corresponding to the models.



(a) MHDs (b) NHDs

Fig. 6. The number of concurrent IP addresses being used over a week (we only plot the results of the five weekdays).

DHCP message from a host that connects to a network is a *discover* message). As expected, we observe that user arrivals follow a weekday pattern, and are affected by class schedules. As an example, Figures 4(a) and (b) plot the number of user arrivals from 9am to 3pm on two Tuesdays (one from the Spring dataset and the other from the Fall dataset) and two Thursdays (again one from each dataset) for MHDs and NHDs, respectively. In the Spring dataset, the number of MHD arrivals in a time window is much smaller than that of NHDs in the same time window. The difference becomes much smaller in the Fall dataset, confirming a much faster adoption speed of MHDs compared to that of NHDs.

B. IP Address Allocation Duration

We observe that distributions of IP address allocation duration are similar over different days. In addition, while IP address allocation durations tend to be longer than session lengths (see Section III), we find that the distribution of IP address allocation duration has similar shape as that of session length. Therefore, for IP addresses allocated to MHDs, we also use two-stage or three-stage hyper-exponential models for IP address allocation duration, and use the methodology in Section IV to obtain the various parameters in the models. Figures 5(a) and (b) plot the empirical distribution along with the results using two-stage and three-stage hyper-exponential models, respectively. We again observe both models provide good match, and three-stage hyper-exponential model slightly outperforms the two-stage model.

C. IP Address Usage

Fig. 6 shows the number of concurrent IP addresses that are being used by MHDs and NHDs at one point of time over the five weekdays of a week. The results are for two weeks, one from the Spring dataset (3/19 to 3/23), and the other from the Fall dataset (9/17 to 9/21). During weekends (not plotted in the figure), the number of concurrent IP addresses that are being used is around 70% less than that on a regular weekday. We find that the number of concurrent IP addresses on the same weekday follows a similar trend for MHDs and NHDs.

As expected, IP address usage follows a diurnal pattern on each day. Furthermore, we observe the largest number of IP addresses being used between 9am and 3pm. In addition, IP address usage on Mondays is similar to that on Wednesdays, and IP address usage on Tuesdays is similar to that on Thursdays. This is because user behaviors are significantly affected by class schedules: Monday and Wednesday have similar class schedules, while Tuesday and Thursday have similar class schedules.

Last, we observe that IP address usage of MHDs increases significantly from March to September; while the increase for NHDs is less significant. This motivates the need to predict IP address usage of MHDs (so that network administrators can take proactive actions to satisfy the fast growth of IP address demand). We next present analytical models for this purpose.

D. Models for IP Address Usage

IP address usage at one point of time is determined by two factors: user arrival process and IP address allocation duration. The models for the latter factor have been described in Section V-B. We next present a model for the former factor, followed by two models for IP address usage.

Let $\lambda(x)$ denote the user arrival rate at time x. Based on measurement results (Fig. 4), we assume that user arrivals follow a Poisson distribution with a constant arrival rate over a short period of time, $[t_i, t_{i+1})$, $i = 0, \ldots, n$, and $t_0 = 0$. Then, $\lambda(x)$ can be defined as

$$\lambda(x) = \begin{cases} \lambda_1 & t_0 \le x < t_1 \\ \lambda_2 & t_1 \le x < t_2 \\ \cdots & \\ \lambda_n & t_{n-1} \le x < t_n \end{cases} \tag{9}$$

For ease of exposition, define $\Lambda(x) = \int_0^x \lambda(t)dt$. That is, $\Lambda(x)$ represents the total number of IP addresses that are allocated before time x. Then for $x \in [t_{k-1}, t_k)$, we have

$$\Lambda(x) = \sum_{i=1}^{k-1} \lambda_i \cdot (t_i - t_{i-1}) + \lambda_k \cdot (x - t_{k-1}). \tag{10}$$

We now present models for IP address usage (or the number of concurrent hosts). Let $N(x)$ denote the number of concurrent hosts at time x. With a slight abuse of notation, in the following, we use $f(x)$ to denote the probability density function for IP address allocation duration, which is represented as (1), where $g(x)$ represents either a two-stage or three-stage hyper-exponential distribution function, and the various parameters are obtained as described in Section V-B. Since the change in the number of concurrent hosts is the difference of arrivals and departures at time x, we have

$$\frac{dN(x)}{dx} = \lambda(x) - \int_0^x \lambda(x-t)f(t)dt, \tag{11}$$

where $\int_0^x \lambda(x-t)f(t)dt$ represents the number of hosts that are leaving at time x. Integrating on both sides of (11) yields

$$\begin{aligned} N(x) &= \int_0^x \lambda(t)dt - \int_0^x \int_0^y \lambda(y-t)f(t)dtdy \\ &= \int_0^x \lambda(t)dt - \int_0^x f(t)dt \int_t^x \lambda(y-t)dy \\ &= \Lambda(x) - \int_0^x f(t)dt\Lambda(x-t) \\ &= \Lambda(x) - \int_0^x f(x-t)\Lambda(t)dt \end{aligned} \tag{12}$$

Substituting (1) into (12), we have

$$\begin{aligned} N(x) = \Lambda(x) - (1-b)\int_0^x g(x-t)\Lambda(t)dt \\ - b\int_0^x \delta(x-t-x_v)\Lambda(t)dt \end{aligned} \tag{13}$$

In (13), the third term on the right hand side can be calculated as

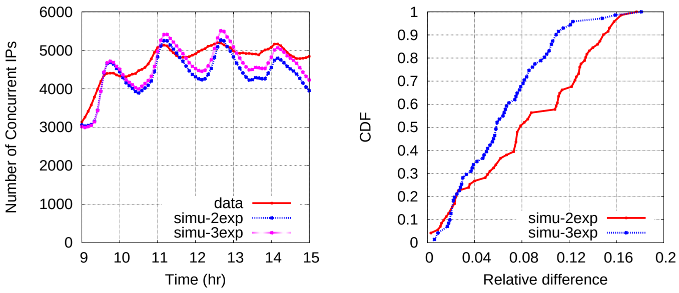
$$b\int_0^x \delta(x-t-x_v)\Lambda(t)dt = \begin{cases} 0, & x < x_v \\ b\Lambda(x-x_v), & x \geq x_v \end{cases} \tag{14}$$

Substituting (10) into the integral part of the second term on the right hand side of (13) yields

$$\begin{aligned} \int_0^x g(x-t)\Lambda(t)dt = \sum_{i=1}^{k-1}\int_{t_{i-1}}^{t_i} g(x-t)\Lambda(t)dt \\ + \int_{t_{k-1}}^x g(x-t)\Lambda(t)dt \end{aligned} \tag{15}$$

For two-stage hyper-exponential model, (15) can be obtained by calculating the integral between $[t_{i-1}, t_i)$ as

$$\begin{aligned} &\int_{t_{j-1}}^{t_j} g(x-t)\Lambda(t)dt \\ &= \int_{t_{j-1}}^{t_j} \mu_1 p e^{-\mu_1(x-t)}(\Lambda(t_{j-1}) + \lambda_j(t-t_{j-1}))dt \\ &+ \int_{t_{j-1}}^{t_j} \mu_2(1-p)e^{-\mu_2(x-t)}(\Lambda(t_{j-1}) + \lambda_j(t-t_{j-1}))dt \end{aligned} \tag{16}$$



(a) (b)

Fig. 7. Validation of the model for IP address usage of MHDs: (a) IP address usage predicted by two-stage and three-stage hyper-exponential models along with the empirical data; (b) the distribution of the prediction errors.

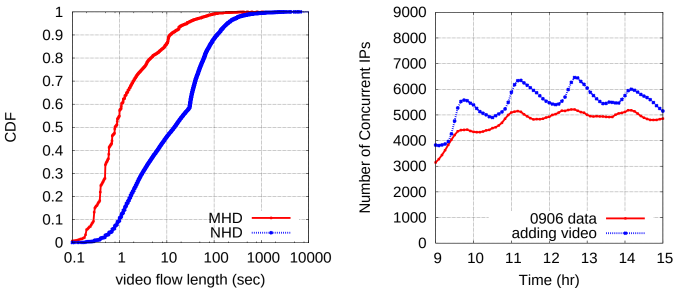
Similarly, for three-stage hyper-exponential model, we have

$$\begin{aligned} &\int_{t_{j-1}}^{t_j} g(x-t)\Lambda(t)dt \\ &= \int_{t_{j-1}}^{t_j} \mu_1 p_1 e^{-\mu_1(x-t)}(\Lambda(t_{j-1}) + \lambda_j(t-t_{j-1}))dt \\ &+ \int_{t_{j-1}}^{t_j} \mu_2 p_2 e^{-\mu_2(x-t)}(\Lambda(t_{j-1}) + \lambda_j(t-t_{j-1}))dt \\ &+ \int_{t_{j-1}}^{t_j} \mu_2 p_3 e^{-\mu_3(x-t)}(\Lambda(t_{j-1}) + \lambda_j(t-t_{j-1}))dt \end{aligned} \tag{17}$$

Combining (13), (14), (15), and (16) or (17), we obtain $N(x)$, the number of concurrent hosts in the network. In this model, we treat $N(0) = 0$. This is because if time 0 is chosen far away from the peak time, then based on session length distribution, hosts that stay in the network at time 0 would have already left the network by the peak time.

E. Model Validation

We now validate the models for IP address usage of MHDs. The models need the IP allocation length distribution and host arrival rate as inputs. As described earlier, IP allocation length distributions for different days are similar, while host arrival rate varies from day to day (See Fig. 4). Hence, to validate the models, we use each day's host arrival rate and the general IP allocation length distribution (i.e., the distribution obtained from all the data) as inputs to the models. As an example, Fig. 7(a) plots the results for MHDs on September 6, 2012. The empirical data, and the predicted results from the two analytical models (by using two-stage and three-stage hyper-exponential models for IP address allocation duration, respectively). We observe a good fit from both models visually. To quantify how well our models predict the number of concurrent IP addresses that are being used, we calculate the prediction error (i.e., the relative difference between a model and the data) every 5 minutes in each day for all the days in the two datasets. Fig. 7(b) plots the CDF (Cumulative Distribution Function) of the prediction errors (we use the absolute values so all the errors are positive). For all the days we examine, the prediction errors are below 15% and 11% for 90% of the days from the two-stage and three-stage hyper-exponential models,



(a) Current video length distribution. (b) Predicted IP address demand.

Fig. 8. Results of the case study: using the analytical models for IP address usage to predict future IP address demand for MHDs.

respectively. This indicates that both models provide accurate prediction, and the prediction from the three-stage model is more accurate.

F. Case Study

Network administrators can use the above analytical models to predict the IP address demand of MHDs when certain change happens in the network. We next present a case study that predicts the IP address demand of MHDs assuming the network usage behavior changes in the future.

To characterize the current network usage behavior, we capture and analyze TCP flows from UConn campus WiFi network over three days (2.9TB of data, see more details in [10]). As an example result, Fig. 8(a) plots the distribution of video lengths for MHDs. For comparison, in the figure, we also plot the distribution of video lengths for NHDs, which tends to be longer (the average length is 53.8s versus 6.6s as in MHDs). We now hypothesize a scenario that assumes that users watch more videos on their MHDs in the future, motivated by the growing popularity of online videos. Specifically, we assume a user watches X more videos in each IP address allocation duration, and X follows a uniform distribution with range of $[0, 300]$. Fig. 8(b) plots the predicted IP address demand by MHDs for this scenario, where the lengths of the videos follow the distribution depicted in Fig. 8(a). We can see that the peak IP address demand will be increased as large as 25% in this scenario.

In the above scenario, we assume the arrival rate of MHDs does not change significantly (specifically, we use the arrival rate of September 6, 2012). Using a similar procedure, we can predict IP address demand of MHDs in a scenario where both the arrival rate and the session lengths are increased, and in many other scenarios.

VI. Related Work

The studies on WiFi networks are extensive, ranging from traffic characterization [33], [11], [19], [32], [24], user activities [34], [16], network performance [6], to mobility modeling [9], [29], [26], [20]. However, most of these studies are for NHDs since MHDs have only been widely adopted recently. In [21], [28], [27], BiPareto distribution is proposed to model the session lengths of wireless devices. Again, the wireless devices are predominantly NHDs. We propose two hyper-exponential models to capture session length distributions of MHDs and find that they provide better fit than BiPareto distribution, probably because the session length distribution of MHDs has a much shorter tail.

Recently, a flurry of studies are on smartphones (e.g., [15], [18], [8], [31], [37], [10], [22], [35]). However, none of them is on characterizing and modeling session lengths and IP address usage of MHDs in WiFi networks as in our study. Our study also differs in scope from several studies on DHCP, which focuses on setting DHCP lease time [25], [30], and debugging DHCP performance [36].

VII. Conclusion and Future Work

In this paper, we have analyzed two five-week long DHCP traces collected from UConn campus WiFi network. We characterized session lengths of MHDs, developed two hyper-exponential models to capture the distributions of session lengths. We further characterized the IP address usage of MHDs, and developed two analytical models to predict IP address usage of MHDs at one point of time, and used a case study to demonstrate the usefulness of our models. Goodness of fit tests indicate that our analytical models of session lengths provide good fit, and evaluation results demonstrate the predictions from our models for IP address usage are accurate.

Our study has been conducted in a specific WiFi network, UConn campus WiFi network. A natural question is whether our results are applicable to other WiFi networks. One of our results demonstrates that, conforming to the nature of MHDs (small pocket devices), MHDs are indeed more mobile and used more opportunistically compared to NHDs, reflected in session lengths and IP address usage. We conjecture that this result might hold in other WiFi networks. As future work, we plan to conduct further studies using data collected from other WiFi networks.

Acknowledgements

We would like to thank J. Farese, S. Maresca, and J. Pufahl of UConn University Information Technology Services for their help with this project. The work was partially supported by National Science Foundation (NSF) under CAREER award 0746841. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors would also like to thank the anonymous reviewers for their insightful comments.

References

[1] Benefits of DHCP. http://technet.microsoft.com/library/cc958943.aspx.

[2] Smartphone Owners Now Outnumber Basic Phone Users, March 2012. http://mashable.com/2012/03/01/smartphones-outnumber-basic-mobile-phone/.

[3] When will smartphones reach saturation in the US?, April 2012. http://www.asymco.com/2012/04/11/when-will-smartphones-reach-saturation-in-the-us/.

[4] Within Five Years, All Phones Will Be Smartphones, April 2012. http://www.slate.com/blogs/moneybox/2012/04/11/all_phones_will_be_smartphones_within_five_years.html.

[5] A. O. Allen. *Probability, Statistics, and Queuing Theory with Computer Science Applications*. Academic Press, 1978.

[6] 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*, Marina Del Rey, CA, June 2002.

[7] A. Balasubramanian, R. Mahajan, and A. Venkataramani. Augmenting mobile 3G using WiFi. In *Proc. of ACM MobiSys*, 2010.

[8] N. Balasubramanian, A. Balasubramanian, and A. Venkataramani. Energy consumption in mobile phones: a measurement study and implications for network applications. In *Proc. of ACM IMC*, 2009.

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

[10] X. Chen, R. Jin, K. Suh, B. Wang, and W. Wei. Network performance of smart mobile handhelds in a university campus WiFi network. In *Proc. of IMC*, 2012.

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

[12] P. Deshpande, X. Hou, and S. R. Das. Performance comparison of 3G and metro-scale WiFi for vehicular network access. In *Proc. of ACM IMC*, 2010.

[13] R. Droms. Dynamic host configuration protocol, 1997. RFC 2131.

[14] H. Falaki and S. Keshav. Trace-based analysis of Wi-Fi scanning strategies. *SIGMOBILE Mob. Comput. Commun. Rev.*, 13(1):73–76, 2009.

[15] H. Falaki, D. Lymberopoulos, R. Mahajan, S. Kandula, and D. Estrin. A first look at traffic on smartphones. In *Proc. of ACM IMC*, 2010.

[16] H. Falaki, R. Mahajan, S. Kandula, D. Lymberopoulos, R. Govindan, and D. Estrin. Diversity in smartphone usage. In *Proc. of ACM MobiSys*, 2010.

[17] A. Feldmann. Fitting mixtures of exponentials to long-tail distributions to analyze network performance models. pages 245–279, 1998.

[18] A. Gember, A. Anand, and A. Akella. A comparative study of handheld and non-handheld traffic in campus WiFi networks. In *Proc. of PAM*, 2011.

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

[20] W. Hsu, T. Spyropoulos, K. Psounis, and A. Helmy. TVC: Modeling spatial and temporal dependencies of user mobility in wireless mobile networks. *IEEE/ACM Transactions on Networking*, 17(5):1564–1577, October 2009.

[21] W.-J. Hsu and A. Helmy. On nodal encounter patterns in Wireless LAN traces. *IEEE Transactions on Mobile Computing*, 2010.

[22] J. Huang, Q. Xu, B. Tiwana, Z. M. Mao, M. Zhang, and P. Bahl. Anatomizing application performance differences on smartphones. In *Proc. of ACM Mobisys*, 2010.

[23] K.-H. Kim, A. W. Min, D. Gupta, P. Mohapatra, and J. P. Singh. Improving energy efficiency of Wi-Fi sensing on smartphones. In *Proc. of IEEE INFOCOM*, 2011.

[24] G. Maier, F. Schneider, and A. Feldmann. A first look at mobile handheld device traffic. In *Proc. of PAM*, 2010.

[25] K. Manas, F. Nick, S. Matt, and C. Russ. Usage-based DHCP lease time optimization. In *Proc. of ACM IMC*, 2007.

[26] X. G. Meng, S. H. Y. Wong, Y. Yuan, and S. Lu. Characterizing flows in large wireless data networks. In *Proc. of ACM MobiCom*, 2004.

[27] 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.

[28] M. Papadopouli, H. Shen, and M. Spanakis. Characterizing the duration and association patterns of wireless access in a campus. *Wireless Conference 2005 - Next Generation Wireless and Mobile Communications and Services (European Wireless), 11th European*, 2005.

[29] M. Papadopouli, H. Shen, and M. Spanakis. Characterizing the duration and association patterns of wireless access in a campus. In *Proc. of European Wireless Conference*, April 2005.

[30] I. Papapanagiotou, E. M. Nahum, and V. Pappas. Configuring DHCP leases in the smartphone era. In *Proc. of IMC*, 2012.

[31] M.-R. Ra, J. Paek, A. Sharma, R. Govindan, M. Krieger, and M. Neely. Energy-delay tradeoffs in smartphone applications. In *Proc. of ACM MobiSys*, 2010.

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

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

[34] I. Trestian, S. Ranjan, A. Kuzmanovic, and A. Nucci. Measuring serendipity: connecting people, locations and interests in a mobile 3G network. In *Proc. of ACM IMC*, 2009.

[35] F. P. Tso, J. Teng, W. Jia, and D. Xuan. Mobility: a double-edged sword for HSPA networks: a large-scale test on Hong Kong mobile HSPA networks. In *Proc. of ACM MobiHoc*, 2010.

[36] B. Vladimir, S. Jesse, and B. Suman. Debugging DHCP performance. In *Proc. of ACM IMC*, 2004.

[37] Z. Zhuang, K.-H. Kim, and J. P. Singh. Improving energy efficiency of location sensing on smartphones. In *Proc. of ACM MobiSys*, 2010.