This paper was presented as part of the main technical program at IEEE INFOCOM 2011

# Modeling and Characterization of Large-Scale Wi-Fi Traffic in Public Hot-Spots

Amitabha Ghosh*, Rittwik Jana†, V. Ramaswami†, Jim Rowland† and N. K. Shankaranarayanan†
*Department of Electrical Engineering, Princeton University, NJ 08544, USA
†AT&T Labs – Research, 180 Park Ave, Florham Park, NJ 07932, USA

***Abstract*—Server side measurements from several Wi-Fi hot-spots deployed in a nationwide network over different types of venues from small coffee shops to large enterprises are used to highlight differences in traffic volumes and patterns. We develop a common modeling framework for the number of simultaneously present customers. Our approach has many novel elements: (a) We combine *statistical clustering* with *Poisson regression* from Generalized Linear Models to fit a *non-stationary Poisson process* to the arrival counts and demonstrate its remarkable accuracy; (b) We model the heavy tailed distribution of connection durations through fitting a *Phase Type* distribution to its logarithm so that not only the tail but also the overall distribution is well matched; (c) We obtain the distribution of the number of simultaneously present customers from an $M_t/G/\infty$ queuing model using a novel *regenerative argument* that is transparent and avoids the customarily made assumption of the queue starting empty at an infinite past; (d) Most importantly, we validate our models by comparison of their predictions and confidence intervals against test data that is not used in fitting the models.**

## I. Introduction

The ubiquitous deployment of wireless hot-spots has drawn millions of users to these networks. These hot-spots provide coverage using the IEEE 802.11 technology by means of Access Points (AP) and are typically managed by Internet Service Providers (ISP) in areas such as coffee shops, book stores, malls, etc. APs are connected to the Internet through a backhaul link, typically a T1 line (see Fig. 1).

While cellular data networks provide a wide area coverage, wireless hot-spots provide a much smaller footprint at considerably higher user bandwidth. Thus, there is reason to believe that users behave differently in cellular and Wi-Fi networks. Since Wi-Fi usage has shown tremendous increase recently, there is a growing need to understand the traffic in large-scale Wi-Fi data networks. This paper is thus a study of public Wi-Fi networks. We demonstrate that there are noticeable differences across different venues for a variety of modeled quantities (e.g., arrival patterns, connection durations, and the number of simultaneous connections). A general modeling framework that can be used for a variety of purposes, however, is then proposed and validated.

Previous studies of workload characterization have been based either on a limited set of users under laboratory conditions or on a short time period [1] over a small number of users typically by active monitoring at venues like technical conferences. Unlike those, our study is based on data from large-scale, publicly deployed Wi-Fi networks. Our goal is to understand typical user workload in a well-provisioned, large-scale environment which includes different types of venues as well as a diverse set of mobile devices. Our data set comes from two large cities in the U.S.

Amitabha Ghosh (Email: amitabhg@princeton.edu) was an intern in AT&T Labs – Research in the summer of 2010 while doing this work.

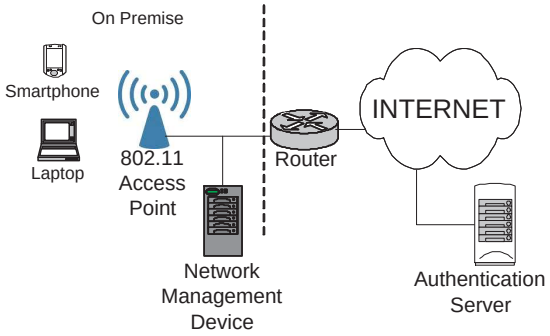



Fig. 1: Mobile Internet access using Wi-Fi hotspots.

We begin by comparing aggregated user workloads both spatially (i.e., across multiple locations for the same business category) and temporally (time of day, and day of week patterns). We group different Wi-Fi business types into a small number of categories: coffee shops and fast food chains; hotels and bookstores; and enterprises and stadiums. These categories are in an increasing order of observed traffic volumes measured in terms of registrations, bytes transferred, etc. For each business type, we construct workload models conditioned on a weekday and a weekend profile. We then compare model predictions to actual measurement traces and validate that our models faithfully capture the aggregate behavior. Our main contributions are:

- *Arrival Patterns:* For all venues under consideration, the arrival patterns show a significant difference between week days and weekends. For instance, during week days in coffee shops, we notice a higher volume marked by three noticeable peaks just before and after normal working hours and during lunch time. As opposed to this, weekends show a smoother process with a single peak in late afternoons in conformity with anticipated human behavior during weekends.
- *Arrival Models:* Although the process of arrivals (registrations at the Wi-Fi AP) differs from venue to venue, it is possible to model these processes using the common model type of *non-stationary Poisson processes*. Further-

more, significant parsimony can be obtained through a common model for sub-clusters of venues of a common type when such venues are clustered by their average volumes. We propose a new approach of combining temporal statistical clustering with the *Poisson regression* technique of the *Generalized Linear Models* (GLM) framework to produce very accurate models.

- *Connection Times:* A key quantity to be modeled is the connection time of a user, which is the interval of time from login to logout at a Wi-Fi port. Since detailed low level measurements at the packet and flow levels are hard to collect and are seldom collected continuously, in practice, many decisions are based on vendor or network provider given capacities for network elements in terms of the number of simultaneously logged-in users they can handle. Our analysis shows the connection distribution to be heavy tailed. The use of a *phase type distribution* for the logarithm of connection times allows us to fit accurately both the tail as well as the head of the distribution. This is important since a large amount of the data (almost $80\%$) tend to be at the head.
- *Simultaneous Users:* We derive the distribution of simultaneously present users based on $M_t/G/\infty$ queues. This distribution is Poisson and matches well with the observed data.

To the best of our knowledge, this is the first time such an approach has been taken to fit a non-stationary Poisson model. Similarly, the modeling of a heavy tailed random variable as the exponent of a *phase type* random variable is new. In addition, this paper distinguishes itself by validating the models using test data not used for modeling.

The remainder of the paper is organized as follows. After reviewing related work in Section II, we provide in Section III a discussion of Wi-Fi arrival patterns, connection durations, and data consumption characteristics by different business types. The details of the models used, the fitting procedures, and a novel proof for a needed key result on $M_t/G/\infty$ queue are presented in Section IV. Section V provides experimental validation of our models by comparing observed traces over the validation period with model based expected values and confidence bounds. Finally, we give some concluding remarks in Section VI.

## II. Related Work

There have been several studies [2], [3] that characterize workload for wireless networks, in particular, Wi-Fi networks. Menasce [4] investigates workload characterization for access to e-commerce sites. He presents a general methodology to classify workload under different functional categories namely, business, session, and requests. Our work looks at multiple business types at a scale in which data is usually collected and made available in large-scale public Wi-Fi hot-spots.

Balachandran *et al.* [1] study wireless user behavior of attendees at a technical conference over three days. The study shows interesting characterizations of Wi-Fi workload; however, it is on a specialized type of users, and arrival counts are modeled as a Markov-modulated Poisson process (MMPP). An MMPP is often preferred for its high versatility in capturing bursty traffic sources (e.g., IP traffic). However, for capturing the almost deterministic, repetitive patterns across days of the week as well as the intra–day patterns, one would require a very large number of phases. Furthermore, the randomness of the phase durations will over time fail to maintain the exact periodicities seen in the actual arrival process. We thus propose a time clustering and a non-homogeneous Poisson process.

Campus *et al.* [5] provide a spatio-temporal model of traffic workload in a campus environment. They also model the session arrival process as a time-varying Poisson process. Our approach takes this one step further by adopting a fitting methodology that is able to reproduce the intra–day patterns accurately. We fit a Poisson regression model to the arrival process and estimate the number of simultaneous ongoing connections using a $M_t/G/\infty$ queue. Most studies to date appear to have focused on devices like laptops [6]. In contrast, our data is primarily from mobile devices such as smart phones, although they do include laptops and other portable devices. In the context of increasing use of smart phones over other smart portable devices, this may be of added interest. Divgi and Chlebus [7] study the traffic characteristics of a commercial hot-spot network in Australia. However, they do not propose a model, and the data characterization is based on combining all venue types together. We consider both venue specific models and a common model for similar venues.

## III. An Overview of Data

In this section, we describe at a high level the data and user behavior characteristics with respect to several metrics, such as (i) number of connections initiated (i.e., number of users registering at the Wi-Fi APs) during weekdays and weekends; (ii) number of bytes uploaded and downloaded; and (iii) connection durations.

### A. Data Collection and Scale

We use data that was collected in the beginning of 2010 for 30 consecutive days spanned across different types of venues (e.g., coffee shops, fast-food chains, bookstores, hotels, enterprises, etc.) at an aggregation point within the infrastructure. Users authenticate with a centralized authentication server upon association with an AP and disconnect when leaving the premise. Each business type has a network management device on the premise that also collects the total number of bytes consumed per user per connection, the device identifier, etc., (see Fig. 1). Note that, although we have the total number of bytes consumed during a connection, we do not have flow-level data (i.e., TCP flows) within a connection that is typically used to explain data burstiness. Such fine grain data are not collected routinely and are available only through specialized studies. In practice, one has to combine a connection level analysis based on field data with intra-connection models based on specialized studies for a detailed low level performance analysis. This paper, however, does not get into such a detailed level and primarily addresses the

| Data Parameters | Total |
|---|---|
| Num. of Customers | 243,742 |
| Num. of Device types | 10 |
| Num. of Connections | 1,322,541 |
| Num. of Cities | 2 (NYC and SF) |
| Num. of Wi-Fi venues | 362 |
| Num. of Zipcodes | 87 |
| Trace duration | 4 weeks |

TABLE I: Data attributes.

modeling of field collected traffic measurements only. Table I lists some attributes of our data that reinforce the scale.

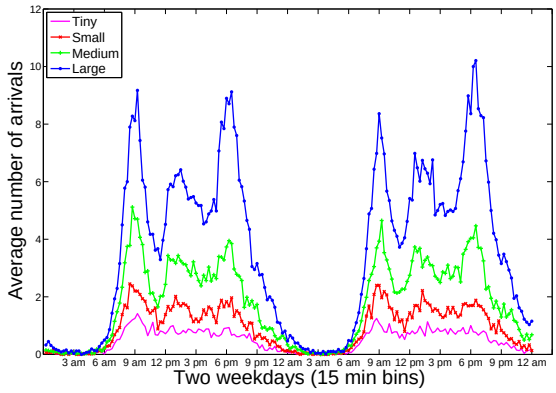
### B. Arrival Distribution

Our preliminary data analysis revealed significant variation in the number of connection arrivals across different venues even within the same business type. To capture such variation, for instance, within the coffee shop venues, we have clustered them into four categories with approximately $60$ venues in each based on the average arrival counts per day. Fig. 2(a) shows a time series plot of this variation in the mean number of arrivals within each group for coffee shops for two weekdays binned every $15$ minutes. Each day starts at 12:00 am (midnight) and ends after 24 hours. We observe three characteristic peaks on both weekdays, which occur typically around 9:00 am, 12:30 pm, and 6:30 pm. This pattern strongly reflects the expected behavior of users having coffee during mornings (possibly before going to work), at lunch times, and after work.

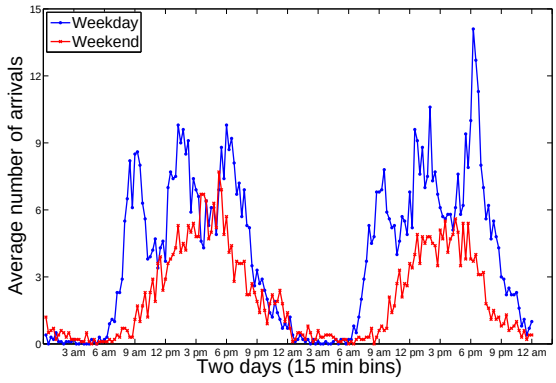
Similar weekday arrival patterns with three characteristic peaks are also observed for bookstores and hotels, as well as for enterprises and stadiums. We show the trends averaged over $20$ book stores and hotels in Fig. 2(b) along with the trend for a typical weekend. We observe significantly different arrival patterns for weekdays and weekends; the latter shows only one peak during the evenings. This trend possibly suggests that customers visit these venues more often during weekend evenings than mornings and noons. Similar weekend arrival patterns are also observed for enterprises; we do not show those graphs here due to lack of space.

### C. Connection Duration Distribution

In this section, we study the connection duration distribution across various business types. A user connection is defined as the time between which a device associates and disassociates with an AP. Fig. 3(a) shows the cumulative distribution function (CDF) of connection durations, and Table II summarizes the means and standard deviations of the connection durations in minutes for the three business types. In Fig. 3(b), we show the complimentary distribution function of connection durations for the three business types. We see that there is a significant number of users with duration less than 20 minutes. We also note that connection durations are heavy tailed and can be expressed using a power law (i.e., for large $x$, the probability of exceeding $x$ is proportional to $x^{-s}$). We observe that in a log-log plot the tail of the distribution has a slope close to $-1$.




(a)




(b)

Fig. 2: (a) Characteristic peaks and variation in average arrival counts across 4 different categories of coffee shops for two weekdays. (b) Average arrival counts across 20 book stores and hotels for two weekdays and two weekend days. Weekdays show 3 characteristic peaks during mornings, noons, and evenings, whereas weekends show only one peak during late afternoons.

| Connection Duration (min) | Mean | s.d. |
|---|---|---|
| Coffee shops and fast food chains | 29.8 | 81.9 |
| Book stores and hotels | 73.4 | 142.3 |
| Enterprises and stadiums | 61.6 | 113.8 |

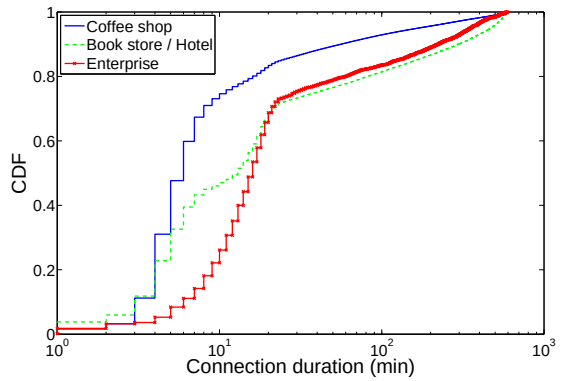
TABLE II: Statistics of connection duration by business types.

### D. Data Volume

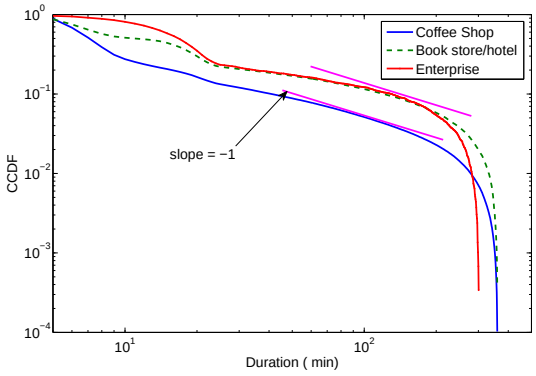
Next we investigate the traffic that is generated by users across different business types. Fig. 4(a) and Fig. 4(b) show the distribution of the downloaded bytes for coffee shop chains and enterprises. For enterprises, we observe that users download a significantly larger number of bytes (peak at 400 KBytes) as opposed to users in coffee shops and fast food chains where the downloaded amount is less than 5 KBytes. There is also an initial peak (at 1 KByte) in both categories attributable to a single TCP packet (MTU of 1500 bytes) for association and immediate disassociation at the APs, as for example, by a moving vehicle in front of a coffee shop.

## IV. Modeling Approach

A complete model of a Wi-Fi system would include models of certain high level entities, such as the process of registrations and logouts, as well as of low level entities like flows and

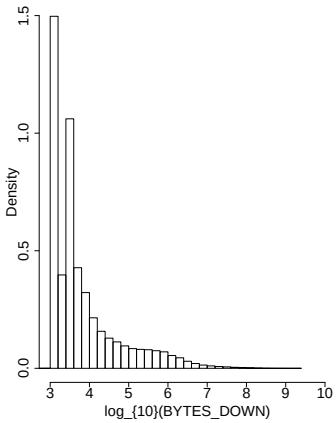



(a)

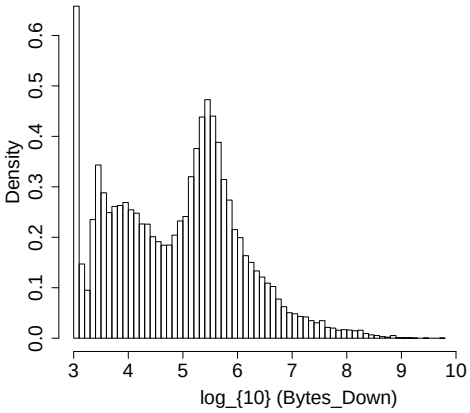



(b)

Fig. 3: (a) CDF of connection duration by business types. (b) Complimentary distribution function of connection durations (log-log scale) by business types.




(a)




(b)

Fig. 4: Downlink byte distribution (a) for coffee shops and fast food chains, and (b) for enterprises.

packet bursts. The low level processes are clearly modulated by the number of *simultaneously present* "connections", which is the number of customers who are logged in at any instant of time. Although the level of congestion and contention for bandwidth depends on the fluctuations at the flow and packet level, obtaining data at such fine granularity is onerous and is seldom made in the field on a continuous basis. Based on some specialized studies using packet level monitoring and deep packet inspections with the specific intent of characterizing detailed behavior of connections, equipment vendors and network providers typically state capacities of network elements like an access point in terms of the number of simultaneous connections that can be supported by them. With this perspective and based on requests from the field, our goal here is to develop a modeling framework to provide reliable predictions for the number of simultaneous connections at a venue. The methodologies, however, apply widely. For instance, we use the ideas in fitting distributions to the connection times to fitting distributions for the bytes downloaded as well.

Besides characterizing existing venues, we want our model to provide a guideline for provisioning new sites as well as to serve as a reliable module generating traffic traces in test environments. The latter requires that our modeling approach be based preferably on a single *parametrized* model class that could be fine tuned to a given venue type and size. One of the major contributions of this work is the demonstration that *non-stationary Poisson processes* provide such a model class and a reliable methodology to fit such a model to actual data. Another contribution is the demonstration that when dealing with a business type that has a very large number of branch locations, we can classify these into a small number of clusters based on the volume and use a common model for each subclass.

### A. Arrival Count Modeling

Notable in the graphs shown in Fig. 2(a) and 2(b) is that the arrival counts show a periodic pattern that repeats over successive weekdays, and within each day there are clearly marked peak periods and low periods. The scaling of the graphs may vary from venue to venue even within the same business type depending on a variety of factors, such as demography, presence of competitors in the neighborhood, etc. In addition, the specific location of the peaks and troughs may depend on the specific venue; thus, while a coffee shop near an office building might see peak traffic during the morning, lunch time, and after work hours of a weekday, another one near a movie theater might experience peaks a while before

and a while after each showing of a movie. Thus, while some customization might be needed to handle such venue specific differences, we wish that a single model type that has a unified methodology and is tractable for further analysis, such as in a queuing model, can faithfully capture the arrival process for all business types. As a final note, we saw that the minor variations between days were not significant for weekdays. Thus, for a given venue, a single model is adequate for all weekdays. However, weekends follow a noticeably different pattern as noted in our discussion of Fig. 2(b).

When dealing with time varying arrival rates, it is customary in the queuing literature to use a *Markov Modulated Poisson Process* (MMPP) [8]. While a MMPP provides a high level of tractability, its Markovian assumptions underlying the constancy periods of arrival rates is a problem. Our data for a given venue exhibit an almost deterministically repeating pattern from day to day, and as can be seen from considering averages across day in Fig. 2(a) an almost deterministic pattern within a day in the mean arrival rates. However, the Poisson structure is well motivated by the fact that the venues have a large number of customers who visit them and only a small fraction of them may indeed use the Wi-Fi facility. Thus, we decided to use a *non-stationary*, *time-dependent Poisson process*.

The usual approaches based on fitting a curve (polynomial or trigonometric) to the observed mean number of arrivals in different periods of the day did not work well due to the problem of not being able to mimic the observed within-day patterns with a small number of terms. Besides, using only the means essentially throws away considerable amount of valuable information available in the detailed data of time stamps of connections from which can get the actual counts for various intervals of time within each day.

Considering the above, we divided the 24 hours of a day into 96 slots of fifteen minutes each, and then classified the 96 slots into a set of 8 clusters using a statistical clustering analysis (k-means clustering [9]). Basically, our clustering analysis clusters time slots into groups such that within each group the average number of arrivals do not differ much. For a set of 238 outlets of a coffee chain, Table III gives a listing of the 96 slots of fifteen minutes starting at 12 AM, and the corresponding clusters into which these are classified. Note the 24 hour wrap around that is achieved automatically by the clustering procedure demonstrated by the fact that the slots just before and after 12 AM get classified into the same cluster. Also, the clustering is based on the statistical patterns and not on the contiguity of the slots; see for example, how the non-contiguous busy slots between 35–37 and 72–75 map to a common cluster although these are not contiguous.

Having assigned time slots to the clusters, for each observation $X$, we can now associate the tag of the cluster ($I = 0, \ldots, 7$) and the tag of the specific slot ($J = 1, 2, \ldots$) to which it corresponds. This allows us to consider these as auxiliary variables and to fit our non-stationary Poisson model using the *Poisson regression* procedure of the *Generalized Linear Models* (GLM) framework used commonly in statistics. The philosophy underlying this method is to model the variable of interest $X$ to be a Poisson random variable with mean $\lambda$ such that $\log(\lambda)$ is a linear function of a set of auxiliary variables, and then to estimate the coefficients in the linear function by a maximum likelihood procedure. In our case, we assume the linear function to be

$$\log(\lambda(I, J)) = \alpha + \sum_{k=1}^{3} \beta_k I^k + \sum_{r=1}^{3} \gamma_r J^r + \delta(I * J). \quad (1)$$

Here $\alpha$ can be considered as characterizing the over–a–day mean behavior, and the other terms as characterizing the differential effects of the specific cluster and slots within it. The above set up uses a third degree polynomial in $I$ and $J$, and through the last term on the right an *interaction* that takes into account the fact that the differential effect of slot $J$ does not have to be the same across all clusters.

A naive Poisson regression without clustering the time slots did not work well. The decision to use a clustering scheme was based on much exploratory data analysis that among others revealed two time scales in operation: one of the order of 1.5 to 3 hours marking sudden rise and fall in the arrival rates, and another of finer fluctuations in counts in intervals of 15 or 30 minutes.

Poisson regression fitting is a standard procedure in most statistical packages, and we use the software R [10] to fit such a model for each venue under consideration. In each case, all model parameters are statistically significant with a high level of significance ($10^{-8}$ and smaller). Due to space limitations, we do not present the detailed R outputs here, but the results of the predictions obtained through our fitting procedures and a validation based on comparison to test data not used to fit the model are presented in Section V.

### *B. Connection Duration Modeling*

With a distribution of the connection times (i.e., duration of the interval between registration and logout), it is possible to build a $M_t/G/\infty$ queuing model to determine the distribution of the number of simultaneous connections present at time $t$. Thus, we now turn to modeling connection durations.

The observed data on durations covered a wide range of connections lasting from a few seconds to several hours with a very long tail; see Fig. 3(b). The empirical tail suggests the use of a a heavy tailed distribution. However, sizable amount of mass is in the lower part of the distribution – for example $78\%$ of the connections are of duration at most 10 minutes.

| Cluster | Slots |
|---|---|
| 0 | 1–28, 94–96 |
| 1 | 83–87 |
| 2 | 32–34, 38–40, 64–66 |
| 3 | 35–37, 72–75 |
| 4 | 48–58 |
| 5 | 41–47, 59–63, 67–71, 76–78 |
| 6 | 79–82 |
| 7 | 29–31, 88–93 |

TABLE III: Clusters of 15 minute time slots over a day.

An accurate determination of the number of simultaneously present customers would thus require a model that not only captures the tail behavior well but also matches well the distribution in its *entire range*. There is also a very small fraction (less than $0.001$) of data with durations less than 1 minute, and we have chosen to ignore them. Similarly, we have ignored a very small fraction of data above 6 hours since this exceeds reasonable validity periods of assigned DHCP authentications and/or corresponds to a small number of "permanent" connections like a store computer. We model the rest as the exponential of a *Phase Type* random variable.

Phase type distributions introduced by Neuts [11], [8] are well-known. Modeling a quantity as a phase type random variable amounts to assuming that it can be decomposed as the sum of a possibly random number of exponential random variables. Such a distribution is characterized as the *absorption time distribution* in a finite $(m+1)$ – state continuous time Markov chain with one absorbing state $m+1$, an initial probability vector $(\alpha, 0)$ and a matrix $T$ of order $m \times m$ that gives the rate of transitions among the first $m$ transient states. Since the rates of absorption are clearly given by the vector $-T\mathbf{1}$, where $\mathbf{1}$ is a column vector of 1's (the zero row sum property of infinitesimal generator of Markov chains), such a phase type distribution is fully characterized by $\alpha$ and $T$ whose order $m$ is also called the order of the phase type distribution. Such a distribution has a density $h(x) = -\alpha e^{Tx} T\mathbf{1}$, and a distribution function $H(x) = 1 - \alpha e^{Tx}\mathbf{1}$. Simple examples of phase type distributions are the exponential distribution and the convolutions and mixtures of finitely many exponential distributions. The phase type class itself is *closed* under the operations of convolutions and mixtures. See cited references in [11], [8] for proofs and more details.

Phase type distributions are also known to be *dense* in the class of all distributions; i.e., they can approximate any distribution on $[0, \infty)$. They have an exponentially decaying tail which asymptotically goes to zero as $e^{-\eta x}$, where $-\eta$ is the real eigen-value of $T$ closest to zero. Assuming that the logarithm of durations follows a phase type distribution permits us to obtain a model that belongs to the class of heavy tailed distributions with decay rate $x^{-\eta}$, while at the same time matches the shape of the empirical distribution across its entire range.

A method for fitting a phase type distribution using the EM algorithm has been provided in [12]. We compute the quantiles of the empirical distribution, form a frequency table based on these, and use the programs provided by the authors of [8]. We have used a set of five phases and found the fit to be very satisfactory.

### C. Simultaneously Present Customers

We assume that the system size or the size of the venue is not a limiting factor. Then, having modeled the arrival (registration) process as a non-homogeneous Poisson process with a time-dependent, deterministic arrival rate $\lambda(t)$, we can, with a model for the distribution of the connection durations (service times), obtain the distribution of the number of simultaneously present connections as that of the number of busy servers in the resulting $M_t/G/\infty$ queue.

We present below a result that derives the necessary time dependent result for the $M_t/G/\infty$ queue. Classical results in the literature of a similar nature [13], [14], [15] use an argument based on Poisson random measures and involve an assumption that the system started empty at an infinite past. Our proof is novel, is based on a simple *regenerative argument* modifying an argument of Ramaswami [8] for the ordinary $M/G/\infty$ model, and avoids the assumption that the emptiness epoch is at an infinite past. With regard to the last, note that in our practical examples, the systems are empty or near empty at midnight when the stores or enterprises are closed, and there is a natural time origin with an empty state for consideration.

**THEOREM** 1: Consider an $M_t/G/\infty$ queue with arrival rate $\lambda(t)$, service time distribution $H(\cdot)$, and starting empty at time $t = 0$. Then the number $\hat{Q}(t)$ of busy servers at time $t$ follows a Poisson distribution with parameter

$$m(t) = \int_0^t \lambda(\tau)[1 - H(t-\tau)]\, d\tau. \tag{2}$$

That is, the probability mass function (pmf) of $\hat{Q}(t)$ is given by

$$\frac{\left[\int_0^t \lambda(\tau)\bar{H}(t-\tau)d\tau\right]^n}{n!} e^{-\int_0^t \lambda(\tau)\bar{H}(t-\tau)\, d\tau}, \tag{3}$$

for $n = 0, 1, \ldots,$ and $t > 0$, where $\bar{H}(x) = 1 - H(x)$.

*Proof:* The number of busy servers at time $t$ is precisely the number $Q(0,t)$ of arrivals in $(0,t]$ that are still in the system at time $t$, and therefore it suffices to show that the probability generating function (pgf) is given by

$$G_t(z,0) = E[z^{Q(0,t)}] = e^{(z-1)\int_0^t \lambda(\tau)\bar{H}(t-\tau)\, d\tau}. \tag{4}$$

The proof becomes easy by an invariant embedding of our problem into the larger problem of determining the distribution of $Q(u,t)$, the number of arrivals in $(u,t]$ that are still in the system at time $t$. To that end, let $G_t(z,u) = E[z^{Q(u,t)}]$ denote the pgf of $Q(u,t)$.

Let

$$\Lambda(t) = \int_0^t \lambda(\tau)\, d\tau \tag{5}$$

denote the expected number of arrivals in $(0,t]$ and in what follows assume, without loss of generality, that the first arrival to our queue will grab a specific server, say server 1, who will not serve any other customers at any later time. This allows us to decouple the behavior of the first arrival from the rest and to use a regenerative argument.

Assume the queue starts empty at $u \le t$. Let $V$ denote the epoch of the first arrival to the queue after epoch $u$. Also, let $W$ be an indicator random variable which takes the value 1 if the first arrival still remains in the system at time $t$, and 0 otherwise. By considering the two cases (i) no arrival occurs in $(u,t]$, and (ii) the first arrival occurs at some $v \in (u,t]$, we

can write

$$G_t(z,u) = E\left[z^{Q(u,t)}\right] = E\left[z^{Q(u,t)}\,\chi(V>t)\right] + E\left[z^{Q(u,t)}\chi(V\le t)\right], \quad (6)$$

where $\chi(\cdot)$ is an indicator function.

Clearly, the first term on the right of (6) evaluates to

$$z^0\,P(\text{no arrival in } (u,t]) = e^{-\int_u^t \lambda(v)\,dv} = e^{-[\Lambda(t)-\Lambda(u)]}. \quad (7)$$

The second term on the right of (6) simplifies to the value

$$\int_u^t \left\{z^0 H(t-v) + z^1[1-H(t-v)]\right\} G_t(z,v)\cdot \lambda(v)e^{-[\Lambda(v)-\Lambda(u)]}\,dv \quad (8)$$

through a regenerative argument whose details are as follows. Given that there is a first arrival in $(u,t]$ at the epoch $v$, one of two things can happen: (a) the arrival at $v$ is gone before time $t$; this is an event with probability $H(t-v)$, and when this event happens, the number $Q(u,t) = Q(v,t)$ is the number of arrivals that come in $(v,t]$ and are still in the system at time $t$; (b) the arrival at $v$ is still in the system at time $t$, which has probability $1-H(t-v)$, and when this happens, $Q(u,t) = 1+Q(v,t)$. Note that, $Q(v,t)$ has pgf $G_t(z,v)$.

Substituting (7) and (8) into (6) and multiplying throughout by $e^{-\Lambda(u)}$ we get

$$e^{-\Lambda(u)}G_t(z,u) = e^{-\Lambda(t)} + \int_u^t \{H(t-v) + z[1-H(t-v)]\} G_t(z,v)\cdot\lambda(v)e^{-\Lambda(v)}\,dv$$

Differentiating the above partially with respect to $u$

$$-\lambda(u)e^{-\Lambda(u)}G_t(z,u) + e^{-\Lambda(u)}\frac{\partial G_t(z,u)}{\partial u} = -e^{-\Lambda(u)}\lambda(u)\{H(t-u) + z[1-H(t-u)]\}G_t(z,u).$$

Rearranging and simplifying,

$$\frac{\partial}{\partial u}G_t(z,u) = -\lambda(u)(z-1)[1-H(t-u)]G_t(z,u). \quad (9)$$

This is a first order ordinary differential equation. Integrating both sides

$$\log G_t(z,u) = -(z-1)\int_0^u \lambda(\tau)[1-H(t-\tau)]\,d\tau + c$$
$$\Rightarrow G_t(z,u) = Ke^{-(z-1)\int_0^u \lambda(\tau)[1-H(t-\tau)]\,d\tau}.$$

Using the obvious boundary condition $G_t(z,t) = 1$, which follows from the fact that $Q(t,t) = 0$, we get

$$K = e^{(z-1)\int_0^t \lambda(\tau)[1-H(t-\tau)]\,d\tau}.$$

Therefore, substituting $K$ and putting $u = 0$ in (5), we get

$$G_t(z,0) = e^{(z-1)\int_0^t \lambda(\tau)\bar{H}(t-\tau)\,d\tau},$$

which is (4), and the theorem follows. ∎

The computation of the integral in (2) is done numerically using a quadrature formula; we use Simpson's 1/3 rule which has an error of the order $O(h^5)$ in the step size. Noting that $\lambda(\cdot)$ is constant over the successive 15 minute intervals, an extremely efficacious programming of the steps of the numerical integration can be made by just writing a function that numerically evaluates the integral

$$\int_a^b [1-H(t-u)]\,du = \int_a^b \alpha e^{T\log(t-u)}\mathbf{1}\,du. \quad (10)$$

Note that once the means in (2) are computed at selected time points, from the associated Poisson distributions, we can construct confidence bounds for the number of simultaneously present customers. Comparison of these with the observed paths of the empirical processes can give yet another validation for the overall modeling procedures. See Section V for such results.

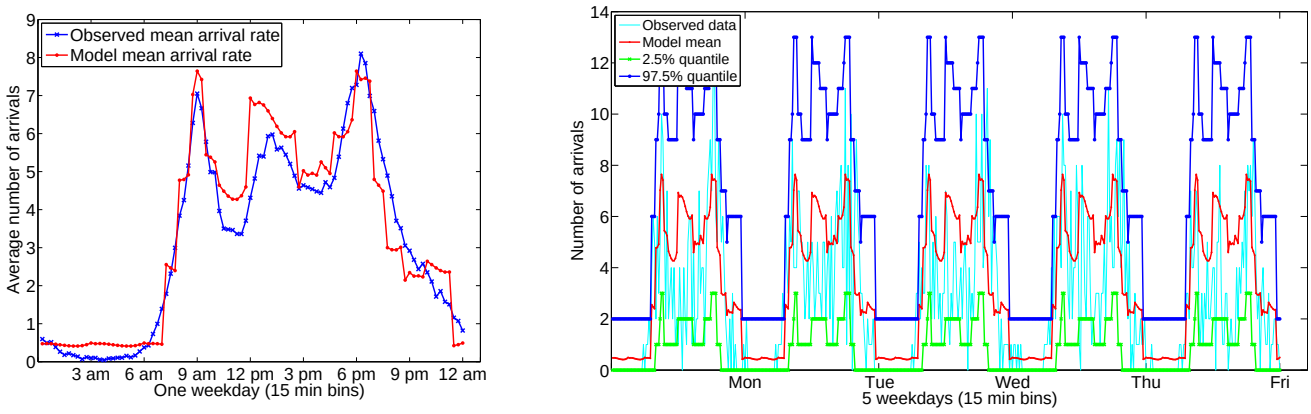
## V. Results and Validation

As noted earlier, our data set comprised of four weeks of records on all Wi-Fi logins in two major U.S. cities in a diverse set of business types. For brevity, we shall in this section confine ourselves to only one of them, namely coffee shops of which we have 238 in our data set. We use the first three weeks of data for modeling and the last week's data for validation. The total number of logins under consideration in this set is 694501 for the learning set of the first three weeks, and 225085 for the test set of the last week. For each connection, we have the login and logout time stamps in addition to the bytes-in and bytes-out counts.

### A. Arrival Process

The starting point of our analysis is the computation of 15 minute counts of arrivals (logins into the Wi-Fi APs). After determining through a preliminary data analysis that the intra-day patterns are reasonably similar across the 238 coffee shops, we use the combined data to obtain a clustering of the time slots into a set of eight clusters such that within each cluster the behavior is similar. The results of the temporal clustering analysis have already been shown in Table III. For each coffee shop in our set, we then fit a non-stationary Poisson process using the Poisson regression method of GLM. This gives us the arrival rates $\lambda(I,J)$ for each time slot for each individual venue.

As noted earlier, it is not efficient to have a separate model for each venue within a business class. However, as the individual branches of the coffee shop chain differ significantly from one another, particularly in the total volume, we cluster the coffee shops into four groups – tiny, small, medium, large – using a statistical clustering analysis ($k$-means clustering) using the observed average counts over the 96 time slots of a day. For each cluster, we average the observed slot arrival counts over the last five weekdays and over all the venues in that cluster to get an average intra–day profile for that cluster, and compare these to the averages of the $\lambda$ values obtained in the Poisson regression now averaged over all members of the cluster. The empirically obtained cluster profiles are very close to the model estimated values, and in Fig. 5(a), we show the




(a) (b)

Fig. 5: Coffee shops: (a) Observed mean arrival rate plotted against the model mean arrival rate; these provide intra–day patterns for a cluster by averaging over its members. (b) Model mean arrival rate along with the $97.5\%$ quantile and $2.5\%$ quantile bands plotted against 5 weekdays of validation data for an example coffee shop.

results for the "Large" cluster as an example. The blue curve corresponds to the empirical means over the last five days and the red curve shows the estimated parameter values averaged over the venues. Thus, if one were to use a representative parameter set for the cluster of large coffee shops, one could indeed use the values given by the red curve in the figure.

We also validate the individual venue models by comparing the observed paths over the last five days to confidence bounds obtained from the non-stationary Poisson model. This also shows that the models perform extremely well. In Fig. 5(b), we show for an example large coffee shop the expected slot means (red curve); the $2.5\%$ (green) and $97.5\%$ (blue) quantiles obtained from the appropriate Poisson distribution; and the actual sample path (cyan) over the five days. The $96 \times 5 = 480$ data points are such that only $6.35\%$ of them falls outside the confidence bounds. The percentage of data points lying outside the confidence band in the weekday data corresponding to the four categories (tiny, small, medium, large) are $1.24\%$, $1.94\%$, $2.86\%$, and $6.35\%$, respectively. The corresponding percentage of those falling outside the confidence interval for the weekend data are $1.55\%$, $2.17\%$, $3.41\%$, and $6.23\%$.

The slightly elevated incidence of points outside the confidence bounds occur in the large category. This could be the result of having a small amount of validation data. There is also a possibility that in the busiest hour, the Poisson model is inadequate and actual arrivals show greater burstiness than what is predicted. This deserves further examination. Development of non-stationary bursty models and their fits using a similar approach as we have taken here is, however, truly an area for further research and beyond our present scope.

### *B. Duration Distribution*

Since the connection durations do not differ significantly across different branches of the coffee shop chain, we have pooled the data together for obtaining a model for the duration $D$ of a connection. As noted earlier, we have eliminated from the data set the very small percentage of connection

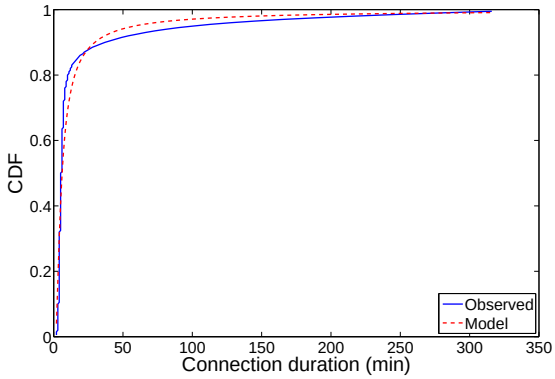



Fig. 6: CDF plot of durations for observed data and model.

durations less than 1 minute and over 6 hours, and have fitted a phase type distribution to $\log(D)$ using the EM algorithm. A phase type fit of order 5 is found to be adequate, and the fitted distribution is characterized by the parameters $\alpha = (1, 0, 0, 0, 0)$ and

$$T = \begin{bmatrix} -\mu & \mu & 0 & 0 & 0 \\ 0 & -\mu & \mu & 0 & 0 \\ 0 & 0 & -\mu & \mu & 0 \\ 0 & 0 & 0 & -\mu & \mu \\ \xi & 0 & 0 & 0 & -\mu \end{bmatrix},$$

where $\mu = 2.747848$ and $\xi = 0.277521$. This distribution is the geometric mixture of the successive convolutions of an Erlang distribution of order 5 with mean $5/\mu = 1.819606$, where the weights of the mixture are given by $(1-\theta)\theta^{n-1}$, $n \geq 1$, where $\theta = \xi/\mu = 0.1001$. It has density

$$f(x) = \sum_{n=1}^{\infty} (1-\theta)\theta^{n-1} \frac{\mu^{5n}}{(5n-1)!} e^{-\mu x} x^{5n-1}, \quad 0 < x < \infty.$$

The way the EM algorithm is designed is that the mean of this model given by $\left(\frac{1}{1-\theta}\right)\left(\frac{5}{\mu}\right) = 2.024$ equals the sample mean

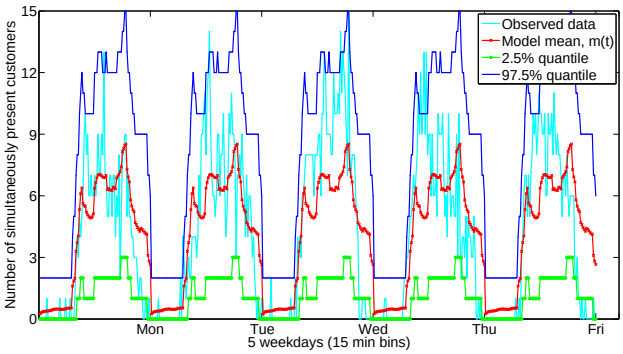



Fig. 7: Expected number of simultaneously present customers along with the $97.5\%$ quantile and $2.5\%$ quantile bands plotted against 5 days of validation data for an example coffee shop.

of $\log(D)$. For the fitted phase type distribution, the eigenvalue of $T$ closest to zero is $-1.01603$, whence the tail of the connection times are $O(x^{-1.01603})$. In Fig. 6, we show a plot of the empirical and fitted cdfs of connection durations, which show a good match across the entire range of the data.

### C. Simultaneously Present Customers

As noted in Section IV-C, we compute the expected number of simultaneously present customers at instants of time by evaluating numerically the integral for $m(t)$. Note that the arrival parameters $\lambda$ computed earlier are 15 minute rates and these need to be divided by 15 to get per minute rates before applying the formulas. Without belaboring much, as one would anticipate from the quality of the fits to the arrival process and the duration distribution, the model does predict the number of simultaneously present customers $Q(t)$ accurately. As a demonstration, for an example coffee shop we show in Fig. 7 $m(t) = E[Q(t)]$ (red curve) after computing it at a grid of values separated by 15 minutes and the confidence intervals at $2.5\%$ (green curve) and $97.5\%$ (blue curve) obtained from the Poisson distributions with parameters $m(\cdot)$. Against these are plotted five days of actual data from the last week. We note that a small percentage of the test data $6.85\%$ falls outside the confidence intervals.

## VI. Conclusions

We examined the characteristics of traffic observed at a large number of public Wi-Fi hot-spots deployed in two large metropolitan cities in terms of arrival counts and temporal variations, connection durations, byte counts, etc. The different venues were categorized into some business types, such as coffee shops and fast food chains, book stores and hotels, and enterprises, and we examined the salient differences among these venue types.

The most notable contributions of this paper are in the area of modeling where we have introduced several novelties. Statistical clustering algorithms and Poisson regression were used to fit non-stationary Poisson models to arrival counts, and models were validated against test data. To the best of our knowledge, this is the first time such an approach has been taken, and the success reported here has opened some interesting possibilities for infusing a greater level of statistical methods in the context of stochastic processes. We also introduced a new way of modeling random variables with long tailed distributions; our choice of modeling their logarithm as a phase type distribution appears to be powerful in that not only the tail behavior but also the overall shape can be matched better. This is yet another area that deserves further theoretical examination. We developed a method for estimating the number of simultaneously present customers using an $M_t/G/\infty$ model and, in that process, discovered a new argument that is direct and avoids unnecessary assumptions about the process.

The practical uses of our models and methods are several: they help in capacity planning for sites wherein decisions are based on vendor or network provider estimates of equipment capacity stated in terms of simultaneously present users due to the inability to collect detailed low level data on a continuous basis; they can be used for building load modules for test purposes; finally, they are useful in the context of network monitoring for detection of changes in traffic patterns, intrusion, etc. In all these contexts, the ability demonstrated here to model diverse venues with a common model class, and to model large sub-clusters of venues within a business type with a common set of parameters, are particularly useful for practical purposes.

## References

[1] A. Balachandran, G. M. Voelker, P. Bahl, and P. V. Rangan, "Characterizing User Behavior and Network Performance in a Public Wireless LAN," in *SIGMETRICS*, June 2002, pp. 195–205.
[2] S. Karpinski, E. M. Belding, and K. C. Almeroth, "Towards Realistic Models of Wireless Workload," in *WiNMee*, Apr 2007.
[3] X. Meng, S. H. Y. Wong, Y. Yuan, and S. Lu, "Characterizing Flows in Large Wireless Data Networks," in *MOBICOM*, Sept 2004, pp. 174–186.
[4] D. Menasce, "Workload Characterization," in *IEEE Internet Computing*, October 2003, pp. 89–92.
[5] F. Hernández-Campos, M. Karaliopoulos, M. Papadopouli, and H. Shen, "Spatio-Temporal Modeling of Traffic Workload in a Campus WLAN," in *WICON*, Aug 2006.
[6] M. Papadopouli, M. Moudastos, and M. Karaliopoulus, "Modeling Roaming in Large-scale Wirless Networks Using Real Measurements," in *WoWMoM*, June 2006, pp. 536–541.
[7] G. Divgi and E. Chlebus, "User and Traffic Characteristics of a Commercial Nationwide Wi-Fi Hotspot Network," in *PIMRC*, Sept 2007, pp. 1–5.
[8] G. Latouche and V. Ramaswami, *Introduction to Matrix Analytic Methods in Stochastic Modeling*. ASA-SIAM Series on Statistics and Applied Probability, 1999.
[9] J. A. Hartigan, *Clustering Algorithms*. John Wiley & Sons, 1975.
[10] GNU-R, http://www.r-project.org.
[11] M. F. Neuts, *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*. Johns Hopkins University Press, 1981.
[12] S. Asmussen, O. Nerman, and M. Olsson, "Fitting Phase-Type Distributions via the EM Algorithm," *Scandinavian Journal of Statistics*, vol. 23, no. 4, pp. 419–441, 1996.
[13] S. G. Eick, W. A. Massey, and W. Whitt, "$M_t/G/\infty$ Queues with Sinusoidal Arrival Rates," *Management Science*, vol. 39, no. 2, pp. 241–252, 1993.
[14] A. Prékopa, "On Secondary Processes Generated by Random Point Distributions of Poisson Type," *Ann. Univ. Sci. Budapest Sectio Math 1*, pp. 153–170, 1958.
[15] S. G. Eick, W. A. Massey, and W. Whitt, "The Physics of the $M_t/G/\infty$ Queue," *Operations Research*, vol. 41, no. 4, pp. 731–742, 1993.