Markovian Modeling of Wireless Trace Data

Jan Kriege
Department of Computer Science,
TU Dortmund
jan.kriege@tu-dortmund.de

Author prepared version of a paper published in

Proceedings of VALUETOOLS 2017: 11th EAI International Conference on Performance Evaluation Methodologies and Tools, December 5–7, 2017, Venice, Italy

EAI

http://doi.acm.org/10.1145/3150928.3150932

Markovian Modeling of Wireless Trace Data

Jan Kriege
Department of Computer Science
TU Dortmund
jan.kriege@tu-dortmund.de

Abstract

With the increased availability of wireless networks, performance evaluation of these networks has become more important in recent years. Adequate models of wireless networks have to account for the user mobility as the movements of the users can have a large influence on the network performance. In many cases data recorded from real networks serves as basis to model the user mobility. Therefore, appropriate distributions have to be found to model characteristics like dwelltimes from the real world data and different general distributions like Lognormal, Weibull or Pareto have been used in the past. In this paper we present an extensive comparison for the fitting quality of those general distributions with Phase-type distributions (PHDs). Our results suggest that in most cases even small PHDs with four or five states yield a better approximation of the real data than the general distributions.

Keywords: Phase-type distributions, Markovian Arrival Processes, Mobility traces

1 Introduction

The adequate and realistic modeling of the traffic load is a crucial step when building stochastic models of computer and communication networks. With the increased availability of mobile devices and the widespread deployment of wireless networks on campuses and public areas as airports or shopping malls the performance evaluation of wireless networks has become more important in recent years [10]. In contrast to the classic wired networks, models of wireless networks have to account for the user mobility, since changing locations of the mobile users can of course have a large influence on the network performance [18]. It is well known that synthetically generated movement patterns, as e.g. generated by Random Waypoint models, of users are unrealistic in many cases [29] and, thus, may lead to wrong assumptions on the performance of the system that is analyzed [22, 28]. Hence, there is a need for traces with real data from wireless networks and a need to identify and model the characteristic properties from those traces. In recent years various mobility traces have been made public [19] and used for the construction of mobility models [16, 21, 28]. However, these models differ largely in what characteristics from the trace they use and in what distributions are applied to model the data. For e.g. the dwelltimes of the users at an access point of the network different distributions have been used in the literature: [29] uses average values, [28] Pareto and [16] Lognormal distributions. Phase-type distributions (PHDs) [23], that have been widely

used in different application areas [7, 6] and have proven to be very flexible and versatile, have not gained much attention in modeling mobility data.

In this paper we systematically model various characteristics from four different mobility traces with Phase-type distribution, assess the quality of the models and compare it with general distributions used in the literature. Since Phase-type distribution can be easily extended into processes that can capture autocorrelation we also investigate the effect of the incorporation of correlation on the models.

The outline of the paper is as follows. Sect. 2 gives an overview of the related work. In Sect. 3 we introduce the basic definitions and notations used later in the paper. Sect. 4 contains the comparison of the different distributions that model characteristics from the mobility traces. The paper ends with the conclusions in Sect. 5.

2 Related Work

The construction of valid mobility and traffic models for wireless networks can be divided into three steps, i.e. the collection of data from a real network, the analysis of this data to identify important characteristics and phenomena and the modeling of this data by mobility or traffic models. These mobility and traffic models require the parametrization of different distributions for values like dwelltimes or interarrival times of packets. In this paper we aim at constructing those distributions for various characteristics obtained from the real data, that can later serve as a basis for mobility and traffic models. There is a huge amount of literature on all the steps mentioned above and therefore the overview presented here can by no means be complete.

In general real data recorded in wireless environments can be divided into coarse-grained and fine-grained data. Coarse-grained data is usually recorded at network components like routers or access points (APs) and therefore does not contain the exact positions of users in the network but only their associations with the access points. In contrast, fine-grained data contains exact locations of the users and therefore has to be recorded on the users' devices. Consequently, it is much easier to record large coarse-grained traces as they can be obtained from the central network components while fine-grained data requires the cooperation of the users. The CRAWDAD archive [19] is probably the largest source for publicly available traces.

One of the first collection of data from a wireless network is presented in [26]. The authors recorded 12 weeks of traffic data at a computer science building at Stanford University and analyzed the active number of users at different APs and sizes of data packets.

Widely used is the coarse-grained trace recorded at Dartmouth college that contains data from more than 550 APs and several thousand users. [18] presents an analysis of the traffic and its hourly or daily distribution. Moreover, the amount of active users is analyzed, but no modeling of the data is performed. In [10] an older trace from the Dartmouth campus is compared with newer data from the same site and changes in the behavior of the users are pointed out. The Dartmouth traces served as basis for several mobility models. In [29] the authors extract average dwelltimes for the buildings and destination probabilities from the trace data and combine it with a map of the real area. In [16] the authors derive information like speed, pause times and transition probabilities from the trace data and use Lognormal distributions to model the pause

times. [21] proposes a model that can capture the interdependence between space and time in the data. The distributions used in this approach are mostly Weibull. A smaller trace with only a few APs recorded at SIGCOMM conference is presented in [2]. The authors analyze the number of users at APs, the session durations and the throughput at the APs. Further traces have been recorded at ETH Zurich [28] and the University of Southern California [13]. In [28] the dwelltimes are modeled using Pareto distributions.

Fine-grained data is in most cases only available for smaller number of users or periods. For example, [14] recorded the contacts between Bluetooth devices handed out to the participants of IEEE Infocom. In [25] fine-grained traces have been collected from five different sites. The authors use a Levy-walk model to capture the mobility. These traces are also used in [9] where mobility from fine-grained data is characterized using a Hidden-Markov-Model.

In summary, for most modeling approaches the authors either used measures directly derived from the data like the mean or general distributions like Weibull, Pareto etc. Phase-type distributions (PHDs) [23] that have been successfully applied to model data from computer networks and other application areas [7] have not gained much attention so far in modeling mobility data. In particular, there is no systematic analysis to assess the ability of PHDs to capture important characteristic from real world mobility data. In contrast, PHDs proved to be superior to general distributions when modeling (un)availability times from distributed systems [6]. A similar comparison for mobility data is performed in this paper.

3 Background and Definitions

3.1 Trace Definitions

As mentioned in Sect. 2 we can distinguish coarse-grained and fine-grained data. The definitions and the contents of the traces differ accordingly. We use the following notation for the trace data.

3.1.1 Mobility Traces

A mobility trace consists of waypoints that describe the movements of a user in a wireless environment. A mobility trace is a sequence of m waypoints $\mathcal{T}_M = (w_1, w_2, \cdots, w_m)$. We use $\mathcal{T}_M^{(c)}$ and $\mathcal{T}_M^{(f)}$ to distinguish coarse-grained and fine-grained traces if necessary. A waypoint is defined as $w_i = (t_i, u_i, l_i)$ where t_i is a timestamp, u_i is the node/user and l_i the location. We assume in the following that the w_i in $\mathcal{T}_M$ are ordered according to their timestamps. The interpretation of the location l_i depends on whether we are dealing with fine-grained or coarse-grained data. In the former case l_i is a tuple (x_i, y_i) with coordinates, in the latter case an identifier of an access point. We use the special location *OFF* to indicate that a node left the area for a certain amount of time. For a mobility trace we define the following sub-traces that only contain waypoints associated with a specific user or access point.

- $\mathcal{T}_M(u) = \{w_i | (w_i \in \mathcal{T}_M) \wedge (u_i = u)\}$: waypoints of user u.
- $\mathcal{T}_M^{(c)}(l) = \left\{w_i | (w_i \in \mathcal{T}_M^{(c)}) \wedge (l_i = l)\right\}$: waypoints at AP l.
- $\mathcal{T}_M^{(c)}(u, l) = \left\{w_i | (w_i \in \mathcal{T}_M^{(c)}) \wedge (u_i = u) \wedge (l_i = l)\right\}$: waypoints of user u at AP l.

For our experiments we define traces that only contain certain values associated with the waypoints. For the dwelltimes occurring in coarse-grained data, i.e. the times a node stays at a location before moving to another location, we have

$$\begin{aligned}
\mathcal{T}^{(c)}_{Dwell}(u,l) &= \left\{d_i = t_{i+1} - t_i | (w_i, w_{i+1} \in {\mathcal{T}_M}^{(c)}(u)) \wedge\ (w_i \in {\mathcal{T}_M}^{(c)}(u,l))\right\} \\
\mathcal{T}^{(c)}_{Dwell}(l) &= \cup_u \mathcal{T}^{(c)}_{Dwell}(u,l) \\
\mathcal{T}^{(c)}_{Dwell}(u) &= \cup_l \mathcal{T}^{(c)}_{Dwell}(u,l) \\
\mathcal{T}^{(c)}_{Dwell} &= \cup_l \mathcal{T}^{(c)}_{Dwell}(l)
\end{aligned}$$

that contains the dwelltimes of user u at location l, the dwelltimes of all users at location l, the dwelltimes of a user at all locations and all dwelltimes observed, respectively.
$\mathcal{T}^{(c)}_{Arr}(l) = \left\{a_i = t_{i+1} - t_i | (w_i, w_{i+1} \in \mathcal{T}^{(c)}_M(l))\right\}$ contains the interarrival times of nodes at AP l and $\mathcal{T}^{(c)}_{Arr}$ denotes the interarrival times of nodes for the whole area, i.e. the time between two waypoints (ordered in time) that indicate an arrival. These waypoints are the first waypoint of each node and all waypoints of a node following a visit to the *OFF* location.

For the fine-grained traces we assume that the current position is measured in fixed time intervals, i.e. $t_i - t_{i-1} = \Delta$ for all i. Then, a measure of interest is the distance covered between two consecutive time points. Note, that this is equivalent to the average speed between two time points. Let,

- $\mathcal{T}^{(f)}_{Dist}(u) = \left\{\sqrt{(x_{i+1} - x_i)^2 (y_{i+1} - y_i)^2} | w_i, w_{i+1} \in \mathcal{T}^{(f)}_M(u)\right\}$ be the distances covered by user u and
- $\mathcal{T}^{(f)}_{Dist} = \cup_u \mathcal{T}^{(f)}_{Dist}(u)$ be the distances of all users.

3.1.2 Traffic Traces

In addition to mobility data one can obtain traffic traces from a (wireless) network. Let $\mathcal{T}_T = (v_1, v_2, \cdots, v_o)$ be a traffic trace with o tuples $v_i = (t'_i, u'_i, l'_i, s_i)$ where t'_i is a timestamp of a data packet, u'_i is the node/user associated with the packet, l'_i the access point that processed the data packet and s_i the packet size. We define the following subtraces.

- $\mathcal{T}_T(u) = \{v_i | (v_i \in \mathcal{T}_T) \wedge (u'_i = u)\}$: packet data of user u.
- $\mathcal{T}^{(c)}_T(l) = \{v_i | (v_i \in \mathcal{T}_T) \wedge (l'_i = l)\}$: packets from AP l.
- $\mathcal{T}_T([t_{start}, t_{end}]) = \{v_i | (v_i \in \mathcal{T}_T) \wedge (t'_i \in [t_{start}, t_{end}])\}$: packets from the time interval $[t_{start}, t_{end}]$.

Combinations of the subtraces like $\mathcal{T}_T(u,l)$ with packets of user u at access point l can easily be defined in a similar way.

For the experiments we use traces with the interevent time between packets ($\mathcal{T}_{PEv}$) and with the traffic amounts associated with waypoints ($\mathcal{T}_{Amt}$), i.e. we define

- $\mathcal{T}_{PEv} = \left\{b_i = t'_{i+1} - t'_i | v_i, v_{i+1} \in \mathcal{T}_T\right\}$: all interevent times.
- $\mathcal{T}_{PEv}(u) = \left\{b_i = t'_{i+1} - t'_i | v_i, v_{i+1} \in \mathcal{T}_T(u)\right\}$: interevent times between packets of user u.
- $\mathcal{T}^{(c)}_{PEv}(l) = \left\{b_i = t'_{i+1} - t'_i | v_i, v_{i+1} \in \mathcal{T}^{(c)}_T(l)\right\}$: interevent times between packets at AP l.

- $\mathcal{T}^{(c)}_{Amt}(u) = \left\{p_{w_i} | w_i \in \mathcal{T}^{(c)}_{M}(u)\right\}$: traffic amounts of user u.
- $\mathcal{T}^{(c)}_{Amt}(l) = \left\{p_{w_i} | w_i \in \mathcal{T}^{(c)}_{M}(l)\right\}$: traffic amounts at AP l.
- $\mathcal{T}^{(c)}_{Amt} = \left\{p_{w_i} | w_i \in \mathcal{T}^{(c)}_{M}\right\}$: all traffic amounts,

where

$$p_{w_i} = \sum_{v_j \in \mathcal{T}_T(u_i, l_i, [t_i, t_i + d_i])} s_j$$

is the traffic amount associated with waypoint w_i, i.e. the amount of traffic generated by user u_i during his stay at location l_i starting at t_i and lasting for the dwelltime d_i.

3.2 Phase-type distributions

Phase-type distributions (PHDs), originally introduced in [23], describe independent and identically distributed random variables as absorption times of a continuous-time Markov chain. A PHD of order n consists of n transient and one absorbing state [7]. It is described by a $n \times n$ subgenerator matrix $\boldsymbol{D}$ that contains the transition rates between the transient states and an initial distribution vector $\boldsymbol{\pi}$. We have that $\boldsymbol{\pi}\mathbf{1} = 1$, $\boldsymbol{D}(i,i) < 0$, $\boldsymbol{D}(i,j) \geq 0, i \neq j$ and $\boldsymbol{D}\mathbf{1} \leq \mathbf{0}$. Furthermore, we denote by $\boldsymbol{d} = -\boldsymbol{D}\mathbf{1}$ the transition rates to the absorbing state.

The behavior of a PHD is as follows: It starts in a transient state according to $\boldsymbol{\pi}$, moves between the states according to $\boldsymbol{D}$ and finally generates an event when the absorbing state is reached.

Properties of the distribution can be expressed in terms $(\boldsymbol{\pi}, \boldsymbol{D})$, i.e. for the moments, probability density function and cumulative distribution function we have

$$\mu_i = E(X^i) = i!\boldsymbol{\pi}\boldsymbol{M}^i\mathbf{1} \tag{1}$$

$$f(x) = \boldsymbol{\pi}e^{\boldsymbol{D}x}\boldsymbol{d}, \quad F(x) = 1 - \boldsymbol{\pi}e^{\boldsymbol{D}x}\mathbf{1} \qquad x \geq 0 \tag{2}$$

where $\boldsymbol{M} = -(\boldsymbol{D})^{-1}$ and $e^{\boldsymbol{D}x}$ is the matrix exponential.

Depending on the structure of $(\boldsymbol{\pi}, \boldsymbol{D})$ several sub-classes can be defined. Well known are the Exponential and Erlang distributions. In this paper Hyper-Erlang distributions (HErDs) [27] are used, that are a mixture of Erlang distributions, i.e. Erlang distribution E_i is taken with probability τ_i.

3.2.1 Parameter Estimation of Phase-type distributions

Parameter estimation or fitting of PHDs denotes the determination of the entries in $(\boldsymbol{\pi}, \boldsymbol{D})$ according to trace data $\mathcal{T}$. Since PHDs are univariate distributions all the derived univariate traces defined above like dwelltimes $\mathcal{T}^{(c)}_{Dwell}$ or time between packets $\mathcal{T}_{PEv}$ are applicable for fitting. Traces like $\mathcal{T}_M$ consisting of tuples can of course not be estimated by a single PHD.

There are basically two classes of approaches to estimate or fit the parameters of a PHD to trace data. Moment based fitting techniques [4, 11, 12] try to find a PHD $(\boldsymbol{\pi}, \boldsymbol{D})$ such that the lower order moments of the PHD (Eq.1) approximate or match the empirical moments of the trace. While methods of this type are very efficient in most cases, maximum likelihood

based approaches are usually slower but provide better results. For a PHD $(\boldsymbol{\pi}, \boldsymbol{D})$ and a trace $\mathcal{T}$ the likelihood is defined as

$$\mathcal{L}_{(\boldsymbol{\pi},\boldsymbol{D})}(\mathcal{T}) = \prod_{x\in\mathcal{T}} f_{(\boldsymbol{\pi},\boldsymbol{D})}(x) = \prod_{x\in\mathcal{T}} \boldsymbol{\pi} e^{\boldsymbol{D}x}\boldsymbol{d}. \tag{3}$$

Since the likelihood only requires the density $f(x)$ it can be defined similarly for any other distribution with given density. Then, we get the maximization problem

$$\mathcal{L}^*(\mathcal{T}) = \max_{(\boldsymbol{\pi},\boldsymbol{D})} \left(\mathcal{L}_{(\boldsymbol{\pi},\boldsymbol{D})}(\mathcal{T})\right). \tag{4}$$

Often the logarithm of the likelihood (log-likelihood) is used, since it does not change the maximum but avoids computation of the product in Eq. 3. Maximum likelihood estimation is done by expectation maximization (EM) algorithms [1, 27] where newer approaches can be applied to large traces and are still quite efficient.
In recent years several tools have been developed that make fitting approaches for PHDs readily available (e.g. [3, 24]).

3.2.2 Measuring the Fitting Quality

For measuring the fitting quality we use analogue measures as in [17, 15, 6] where a similar comparison was performed for availability and unavailability data from distributed systems. In particular, we compare the quality according to the likelihood that uses the density function and according to the the Kolmogorov-Smirnov (KS) test that uses the distribution function.

The likelihood is a relative measure of the fitting quality, i.e. a single value for a distribution and a trace does not provide any information about the quality of the fitting. But it can be used to compare the quality of two distributions fitted to a trace as the distribution with the larger likelihood value provides a better quality. Thus, we can use likelihood values to compare PHDs of different sizes and to compare PHDs with general distributions like Weibull or Lognormal.

Another measure for the quality of the parameter estimation are statistical tests like KS and Anderson-Darling (AD) test that are for example used in [17, 15] to compare the fitting quality for failure traces. The AD test is, to the best of the author's knowledge, not available for PHDs and therefore we are limited to the KS test in this work that requires the distribution function from Eq. 2.

The KS test returns a p-value and for a p-value below the significance level the hypothesis that the trace is drawn from the distribution should be rejected. For long traces the test will usually result in a low p-value, which is known to be a general problem of goodness-of- fit tests. Therefore, as proposed in [17, 15] we draw 30 samples from the trace randomly and compute the corresponding p-values. This is repeated 1000 times and the average p-value is used as p-value for the distribution.

4 Modeling Wireless Trace Data with Phase-Type Distributions

4.1 Experiment Setup

For our experimental comparison four different data sets from the CRAWDAD archive [19] were used, three of these data sets contained coarse-grained data, one set fine-grained data. Details on the data sets are summarized in Sect. 4.2.

From the data we generated traces defined in Sect. 3.1 and fitted them with five general distributions often used in stochastic modeling (Exponential, Lognormal, Weibull, Gamma and Pareto) using the *fitdistrplus* package for the software R [8] and the SSJ library [20]. For comparison we fitted HErDs with different number of states using the approach from [27]. The fitting quality is measured as described in Sect. 3.2.2.

We ignored very small traces with 5 or less entries. Thus, the number of traces for locations or users reported below might be slightly less in this paper than the numbers reported by the authors that collected these traces. For fitting we scaled all traces to have mean 1.0 to avoid numerical difficulties with very large values when fitting. This is not a real restriction as the matrices of PHDs can be scaled reversely to match the original trace.

4.2 Traces and Trace Preparation

In the following we briefly introduce the data sets from the CRAWDAD archive [19] used in our experiments and explain how they were preprocessed to obtain the trace formats described in Sect. 3.1.

4.2.1 Dartmouth Trace

The coarse-grained data recorded at Dartmouth campus [10, 18] covers user mobility for over three years. The data is divided into several data sets. The movement data contains timestamp and access point associations of all nodes present during the time period. The special access point *OFF* is already included in the data and was used when an access point deauthenticated a node due to inactivity (which happens when the node showed no activity for 30 minutes). Thus, the waypoints $w_i = (t_i, n_i, l_i)$ for $\mathcal{T}_M^{(c)}$ can be easily obtained from the raw data. Additionally, tcpdump data is available from the campus that includes all packets processed by the access points. The traffic trace $\mathcal{T}_T$ with its tuples $v_i = (t'_i, n'_i, l'_i, s_i)$ is generated from this data. Unfortunately the tcpdump data is not available for all access points and, furthermore, the time span for movement and tcpdump data is not identical, such that we do not have traffic information for all waypoints in the trace.

Access points in the trace are denoted as *BuildingType BuildingNumber* AP *APnumber*, e.g. *AcadBldg1AP1* for the first access point in the first academical building. Therefore it is easy to aggregate all APs of the same building into a single location and consider both, individual APs and buildings, as locations for the experiments. If the aggregation results in two consecutive waypoints of a node to have the same location, the two waypoints are joined into a single one.

4.2.2 Stanford Trace

The trace recorded at Stanford's CS Department [26] is another coarse-grained trace that contains data for individual packets from almost three month consisting of timestamp, packet size, corresponding node and the access point used for packet transmission. Thus, the traffic trace $\mathcal{T}_T$ can be obtained immediately from the recorded data. To generate the mobility trace $\mathcal{T}_M^{(c)}$ we aggregated all consecutive packets of a user associated with the same access point into a single waypoint, i.e. the timestamp of the first packet is the arrival time at the access point and the dwelltime is the amount of time between the arrival time and the first timestamp associated with another access point. If there was no activity (i.e. no packet sent or

received) for at least 30 minutes, we assumed that the user logged off (left the area) and the dwelltime only covers the time between arrival time and logoff time. Technically, we introduced an artificial access point denoted as *OFF* that a node visits during long periods of inactivity. This is similar to the Dartmouth trace where the *OFF* location is already included in the original trace.

4.2.3 USC Trace

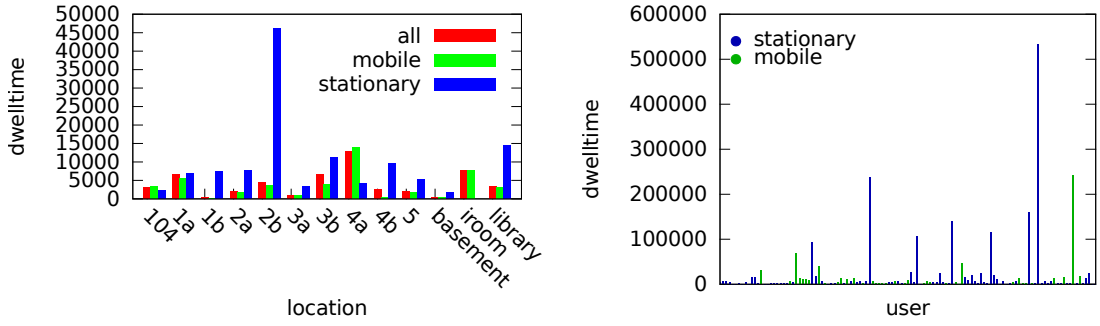
The third coarse-grained trace was collected at University of Southern California [13]. It contains the AP associations from more than a year of measurements and several thousand users. No traffic data is available from the dataset, thus we only have trace $\mathcal{T}_M^{(c)}$.

4.2.4 NCSU

The NCSU data set [25] consists of five data sets with fine-grained data recorded at two university campuses (NCSU and KAIST), New York City, Disney World Orlando and the North Carolina state fair. Each data set contains the positions of users recorded every 30 seconds via GPS resulting in a mobility trace $\mathcal{T}_M^{(f)}$. Traffic information is not available for this data set.

4.3 Interarrival Times and Dwelltimes

We present results for the interarrival and dwelltimes of the coarse-grained traces first. Modeling of the dwelltimes can be done location-based or user-based. Fig. 1 shows the average dwelltimes for different locations and different users from the



(a) APs (Traces $\mathcal{T}_{Dwell}^{(c)}(l)$)

(b) Users (Traces $\mathcal{T}_{Dwell}^{(c)}(u)$)

Figure 1: Stanford: Average dwelltimes

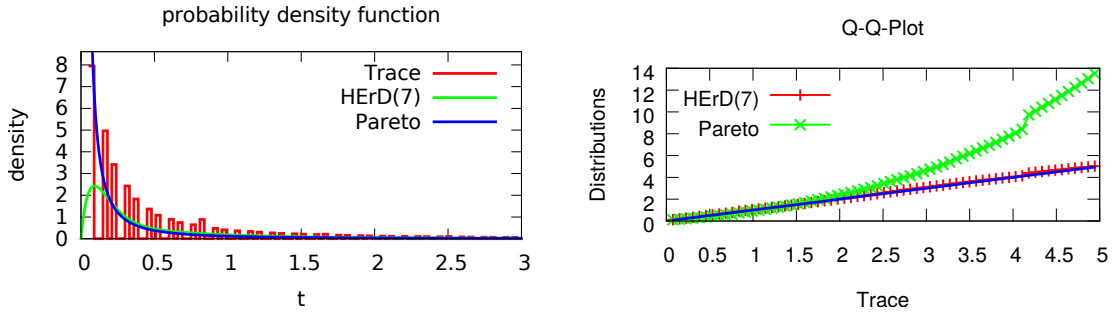
Stanford data. We distinguish mobile and stationary users in the plot. We required a mobile user to have visited at least 3 distinct locations during the recorded time (otherwise the user is a stationary user). As one can see there is a large variation in the average dwelltimes at different access points and for different users. Similar behavior could be observed for the other coarse-grained traces mentioned above.

Due to the sheer amount of traces resulting from the data we can only present detailed results for a few representative traces and aggregate the results for the remaining traces.

Tables 1 and 2 show the log-likelihood and p-values for three traces from the Dartmouth data. In particular, they show

Table 1: Dartmouth: Log-likelihood and p-values for dwelltimes

	$\mathcal{T}_{Dwell}$		$\mathcal{T}_{Dwell}(AcadBldg16)$		$\mathcal{T}_{Dwell}(LibBldg2AP13)$	
	log-likelihood	p-value	log-likelihood	p-value	log-likelihood	p-value
Exponential	-32475499.00	0.00	-68904.00	0.00	-27161.00	0.00
Lognormal	42346574.77	0.46	75082.29	0.21	20461.24	0.46
Weibull	37584020.77	0.29	73783.26	0.32	17588.63	0.33
Gamma	28744307.92	0.06	61581.64	0.25	13242.92	0.11
Pareto	42516187.45	0.25	66750.71	0.14	20367.85	0.22
HErD(2)	33716302.01	0.22	47797.89	0.06	17133.35	0.30
HErD(3)	40960624.28	0.43	**80740.98**	**0.39**	**20509.20**	**0.51**
HErD(4)	**42860557.65**	**0.48**	86304.10	0.49	21585.35	0.55
HErD(5)	43354838.92	0.51	87689.72	0.50	21934.33	0.55
HErD(7)	43605073.87	0.52	88082.13	0.50	21976.01	0.55
HErD(10)	43670471.93	0.51	88147.17	0.50	22276.96	0.54
HErD(15)	44653555.78	0.52	89851.92	0.50	22822.49	0.55



(a) pdf

(b) Q-Q plot

Figure 2: Dartmouth: Results for the interarrival times $\mathcal{T}_{Arr}$ and a fitted $HErD(7)$ and a Pareto distribution.

the results for the interarrival times and dwelltimes from the complete data set, from a building and from an access point. The smallest HErD that yields a better result than all general distributions is printed in bold. As we can see from the results a PHD with three or four states usually yields a better fitting quality than the general distributions. An exception that needs some further explanation is the trace with interarrival times for the whole area where the Pareto distribution has the best likelihood but a very poor p-value. Fig. 2 shows the density functions and a Q-Q plot of the trace, a HErD and the Pareto

Table 2: Dartmouth: Log-likelihood and p-values for interarrival times

	$\mathcal{T}_{Arr}$		$\mathcal{T}_{Arr}(AcadBldg16)$		$\mathcal{T}_{Arr}(LibBldg2AP13)$	
	log-likelihood	p-value	log-likelihood	p-value	log-likelihood	p-value
Exponential	-7731048.00	0.03	-68777.00	0.19	-26867.00	0.00
Lognormal	-4306176.80	0.34	-57242.57	0.35	21101.88	0.52
Weibull	-5768921.22	0.12	-54748.07	0.44	18083.53	0.41
Gamma	-6682602.32	0.12	-58208.07	0.41	10366.29	0.07
Pareto	-2264729.30	0.06	-115150.76	0.00	10070.49	0.02
HErD(2)	-4872580.20	0.29	-55327.98	0.30	15055.50	0.23
HErD(3)	-4654346.94	0.26	**-50780.78**	**0.48**	19374.40	0.44
HErD(4)	-4205057.92	**0.34**	-50681.08	0.48	20753.41	0.51
HErD(5)	-4170095.17	0.35	-50530.02	0.48	20914.91	**0.52**
HErD(7)	-4067847.37	0.35	-50058.19	0.49	20918.20	0.52
HErD(10)	-3871630.79	0.33	-49747.91	0.49	**21290.73**	0.52
HErD(15)	-3755418.86	0.32	-49247.77	0.49	21525.73	0.52

distribution. As one can see, the Pareto distribution has a very high density for the smaller trace values resulting in a large likelihood value, though the long tail of the distribution overestimates the larger trace values dramatically as shown by the Q-Q plot, resulting in the poor p-value as the KS test compares the distribution function.

This behavior could also be observed for other traces where the Pareto distribution resulted in a high likelihood value but a low p-value.

Tables 3 and 4 show results for the complete Stanford data and two APs from the data set that confirm our observations from the Dartmouth traces that in most cases 4 states are sufficient to yield better results than with general distributions. The results for the third coarse-grained trace are shown in Table 5 and fit into the pattern observed so far.

Fig. 3 shows aggregated results for all locations from the Dartmouth data set. We fitted all traces with dwelltimes and interarrival times for APs and buildings with the five general distributions and HErDs with different numbers of states. For each HErD we counted the number of traces where the fitting quality according to log-likelihood and p-value was better than the quality of all general distributions. As we can see from Fig. 3 the HErDs performed better according to the p-values than the likelihood values. This is again due to the Pareto distribution and the phenomenon described above. In general, a $HErD(5)$ was sufficient to obtain the best results for most traces.

Fig. 4 shows the results for a user-based modeling where we generated traces with dwelltimes for every user. Due to the sheer amount of traces we omitted the computation of the p-values here and only present the likelihood values that confirm our observations from the location-based modeling.

Fig. 5 presents the aggregated results for the Stanford data set. Figs. 5a and 5d show the number of traces where the fitting quality of the HErDs according to log-likelihood and p-value was better than the quality of all general distributions.

Table 3: Stanford: Log-likelihood and p-values for dwelltimes

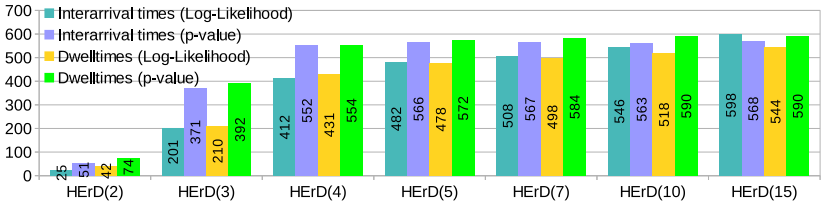
	$\mathcal{T}_{Dwell}$		$\mathcal{T}_{Dwell}(basement)$		$\mathcal{T}_{Dwell}(library)$	
	log-likelihood	p-value	log-likelihood	p-value	log-likelihood	p-value
Exponential	-14135.00	0.00	-304.00	0.00	-1295.00	0.00
Lognormal	5596.64	0.35	124.32	0.44	-388.36	0.44
Weibull	5910.92	0.46	97.49	0.39	-524.84	0.25
Gamma	4066.01	0.32	53.22	0.16	-722.00	0.11
Pareto	3076.15	0.06	162.93	0.10	-1535.03	0.00
HErD(2)	-579.32	0.14	47.75	0.16	-415.10	0.37
HErD(3)	5481.36	0.31	138.36	**0.48**	**-370.63**	**0.46**
HErD(4)	**6966.15**	**0.49**	141.73	0.49	-369.55	0.46
HErD(5)	7260.97	0.49	142.64	0.49	-340.95	0.46
HErD(7)	7371.55	0.49	158.89	0.49	-310.63	0.49
HErD(10)	7411.44	0.49	161.12	0.49	-286.46	0.49
HErD(15)	7903.60	0.49	**169.08**	0.48	-282.84	0.49

Table 4: Stanford: Log-likelihood and p-values for interarrival times

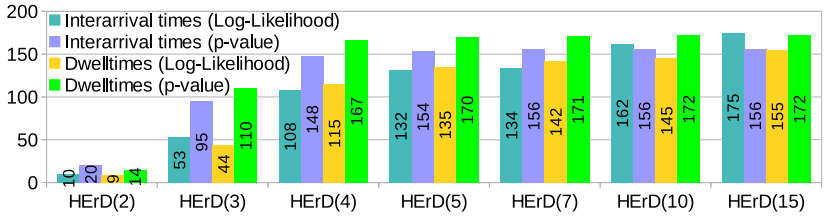
	$\mathcal{T}_{Arr}$		$\mathcal{T}_{Arr}(basement)$		$\mathcal{T}_{Arr}(library)$	
	log-likelihood	p-value	log-likelihood	p-value	log-likelihood	p-value
Exponential	-4028.00	0.11	-302.00	0.00	-1287.00	0.01
Lognormal	-2761.10	0.15	1092.28	0.46	-538.95	0.26
Weibull	-2181.52	0.30	1021.68	0.12	-430.47	0.39
Gamma	-2026.95	0.39	888.91	0.00	-537.11	0.26
Pareto	-3951.14	0.02	1140.19	0.18	-1293.18	0.00
HErD(2)	**-1718.74**	**0.45**	984.55	0.06	-603.92	**0.41**
HErD(3)	-1572.03	0.48	1089.12	**0.47**	-546.11	0.43
HErD(4)	-1481.75	0.51	1122.09	0.48	**-324.08**	0.53
HErD(5)	-1461.31	0.53	1129.27	0.54	-324.08	0.53
HErD(7)	-1458.61	0.53	1130.62	0.54	-322.14	0.53
HErD(10)	-1377.26	0.51	**1141.35**	0.51	-304.03	0.53
HErD(15)	-1307.52	0.52	1146.94	0.51	-278.91	0.53

Table 5: USC: Log-likelihood values for dwelltimes and interarrival times from two APs and a user

	$\mathcal{T}_{Arr}(172.16.8.245)$	$\mathcal{T}_{Dwell}(172.16.8.245)$	$\mathcal{T}_{Arr}(172.16.8.241)$	$\mathcal{T}_{Dwell}(172.16.8.241)$	$\mathcal{T}_{Dwell}(u10100)$
Exponential	-33883.00	-34317.00	-4918.00	-4936.00	-1093.00
Lognormal	-7279.12	11344.06	-1221.28	1820.70	-741.38
Weibull	-11811.81	12097.72	-1631.84	1800.11	-939.48
Gamma	-19248.51	6247.43	-2328.15	325.32	-1030.94
Pareto	-23098.90	-894.58	-4472.71	-1718.86	-1919.40
HErD(2)	-9846.89	-2471.69	-1740.76	1254.86	-773.91
HErD(3)	-7428.95	**14821.07**	-1256.96	**2031.57**	-764.86
HErD(4)	**-7136.56**	16301.54	**-1193.57**	2627.21	**-716.32**
HErD(5)	-7107.49	16502.23	-1193.30	2627.37	-707.99
HErD(7)	-7106.46	16603.97	-1192.45	2723.08	-704.69
HErD(10)	-6966.59	16980.80	-1190.97	2774.99	-689.19
HErD(15)	-6862.18	17139.23	-1174.57	2784.59	-679.39



(a) Access points (598 traces)



(b) Buildings (176 traces)

Figure 3: Dartmouth: Results for dwelltimes (traces $\mathcal{T}^{(c)}_{Dwell}(l)$) and interarrival times (traces $\mathcal{T}^{(c)}_{Arr}(l)$). Number of traces where the $HErD(n)$ is better than general distributions according to log-likelihood and p-values.

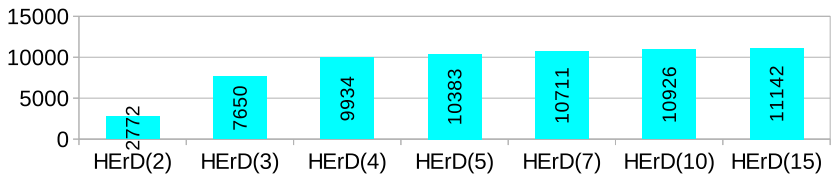
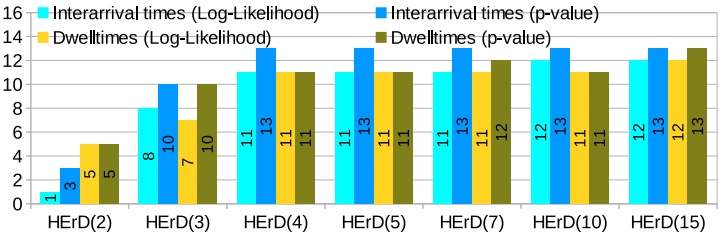
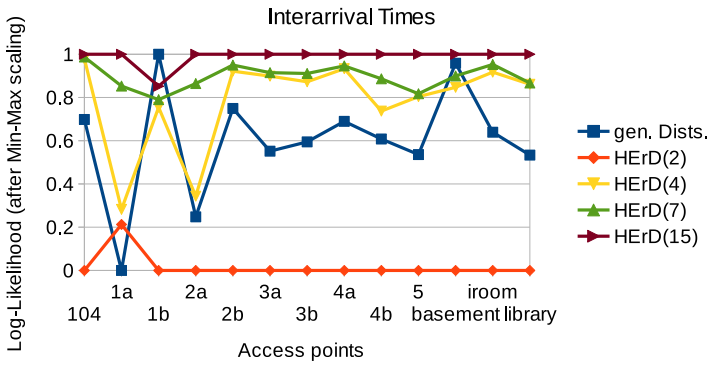


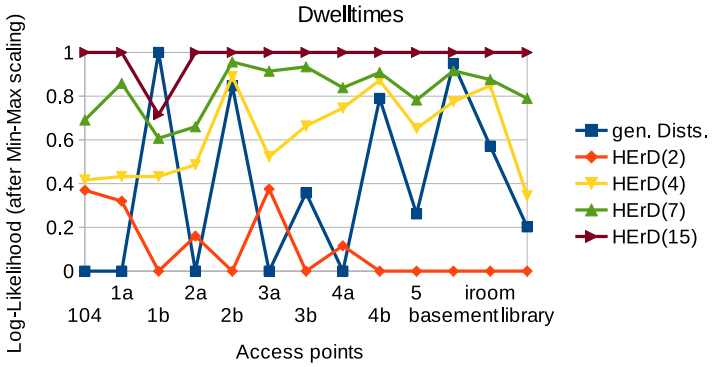
Figure 4: Dartmouth: Results for dwelltimes of users (traces $\mathcal{T}^{(c)}_{Dwell}(u)$, $u = 1, \ldots, 11766$). Number of traces where the $HErD(n)$ is better than general distributions according to log-likelihood.



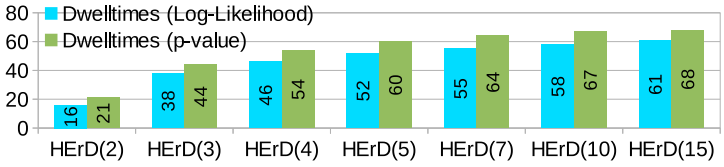
(a) Interarrival and dwelltimes for APs (13 traces)



(b) Interarrival times for APs (13 traces)



(c) Dwelltimes for APs (13 traces)

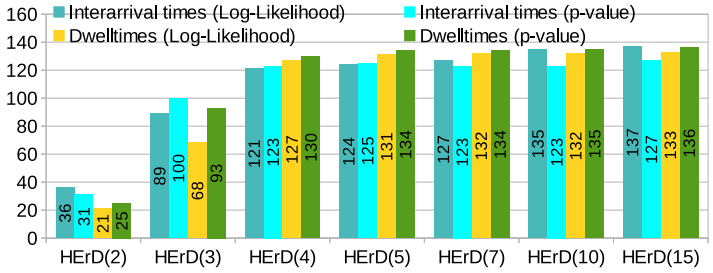


(d) Dwelltimes for users (74 traces)

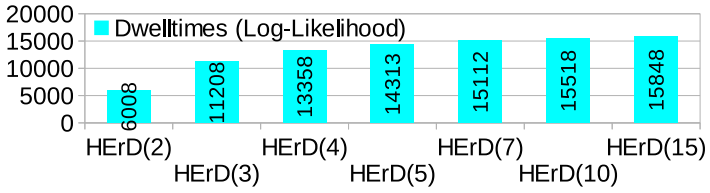
Figure 5: Stanford: Results for dwelltimes and interarrival times. (5a), (5d) Number of traces where a PHD with n states is superior according to log-likelihood and p-values. (5b), (5c) Log-likelihood values (after Min-Max-scaling)

Again, we can conclude that for the traces from the APs in most cases HErDs with 4 states were sufficient. The dwelltimes of the users were more difficult to fit with the HErDs, though for a large majority of the traces they still yielded the best results. For the plots in Figs. 5b and 5c we applied Min-Max scaling to the likelihood values of the distributions for the interarrival and dwelltimes, i.e. we normalized the likelihood values to the interval $[0, 1]$. On the x-axis the diagrams show the 13 APs from the trace and on the y-axis the scaled likelihood values for the best general distribution and four HErDs. With the exception of the APs *1b* and *basement* four states were sufficient for the best results.

Finally, the results for the USC data set is shown in Fig. 6. Again, we omitted the computation of the p-values for the large amount of traces from the dwelltimes for users.



(a) Interarrival and dwelltimes for APs (137 traces)



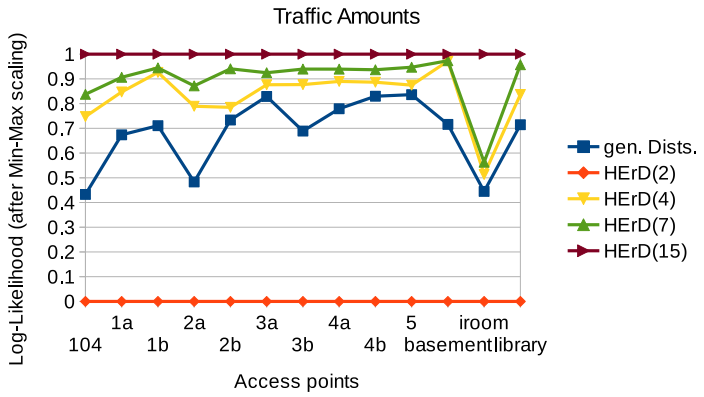
(b) Dwelltimes for users (17078 traces)

Figure 6: USC: Results for dwelltimes and interarrival times. Number of traces where a PHD with n states is superior according to log-likelihood (and p-values).

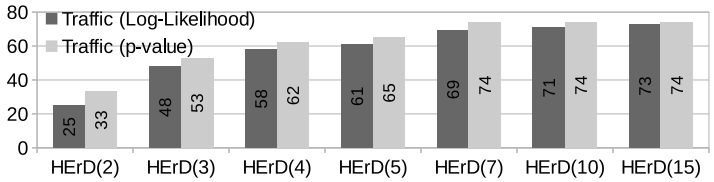
4.4 Traffic

We treated two types of traces regarding the traffic generated by the users. Traces $\mathcal{T}_{PEv}$ consider the times between individual packets, while traces $\mathcal{T}_{Amt}$ contain more abstract data and consider the amount of traffic associated with a waypoint. Traffic data is only available for the Stanford and Dartmouth data sets with some additional limitations. The Dartmouth data only contains packets information for some buildings, thus we have less traces than for the dwelltimes. Timestamps from the Stanford data are given in seconds only, which means that the traces with packet interevent times contain only few distinct and discrete values, which makes them difficult to fit for continuous distributions. Hence, we omitted those traces.

Fig. 7 shows the results for the traffic amounts associated with the different waypoints for the Stanford data. For the traffic amounts at the different APs shown in Fig. 7a we see that with 4 states HErDs yield better results than the general



(a) Access points (13 traces)



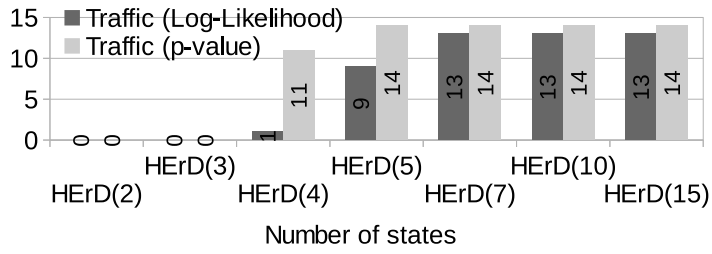
(b) User (74 traces)

Figure 7: Stanford: Results for traffic amounts $\mathcal{T}_{Amt}$ during stay at a location. (7a) Log-likelihood values (after Min-Max-scaling). (7b) Number of traces where a PHD with n states is superior according to log-likelihood and p-values.

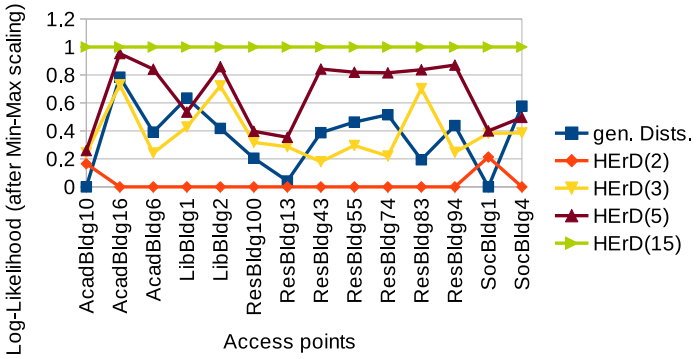
distributions. If we consider traces for the users, fitting was more difficult, although with 4 states the majority of traces was adequately captured by the HErDs. The traffic amounts from the buildings of the Dartmouth data set were more difficult to fit for HErDs as shown in Fig. 8a. It required 7 states to get better results than the general distributions for most of the traces. Fig. 8b shows the scaled likelihood values for the interevent times of packets at different APs. As we can see, 5 states are sufficient for almost all traces here.

4.5 Distances

So far we only presented results for coarse-grained traces. For fine-grained data we use the distances covered between waypoints $\mathcal{T}^{(f)}_{Dist}(u)$ as a measure for comparison. The NCSU data set contains traces from five different sites. Since the number of users differ, we show the percentage of traces where the HErDs resulted in a better approximation than the general distribution in Fig. 9. With 4 states the HErDs are best for more than 80% of the traces from all sites with the exception of one p-value.



(a) Traffic amounts for buildings (14 traces)



(b) Packet interevent times for buildings (14 traces)

Figure 8: Dartmouth: Results for traffic traces from buildings. (8a) Number of traces with traffic amounts ($\mathcal{T}_{Amt}$) where a PHD with n states is superior according to log-likelihood and p-values. (8b) Log-likelihood values for traces with packet interevent times ($\mathcal{T}_{PEv}$).

4.6 Adding Correlation

In the previous experiments we have only fitted the empirical distribution of the trace. However, it is well known that data from real networks might exhibit autocorrelation and dependencies. Fig. 10a shows the first 10 lags of autocorrelation for some of the traces with interarrival times of users at APs. As one can see, the autocorrelation values are significant and in the following we demonstrate that the incorporation of these values will significantly improve the fitting quality. Phase-type distributions can be easily extended into Markovian Arrival Processes (MAPs) [7] that can capture this autocorrelation. We used the fitted HErDs as input to the EM algorithm for MAPs from [5], which implies that the algorithm might change the matrix parameters of the distribution but keeps the structure of the HErD, and performed 10 iterations of the algorithm. Since MAP fitting is much more time consuming and the required time depends on the number of states we only used smaller HErDs with up to 5 states as input. Since the MAPs are also created with an EM algorithm, the likelihood of the HErDs is always at least slightly improved by the MAPs. However, the comparison of the likelihood values in Fig. 10b shows that by using a MAP with 4 or 5 states we obtain a better fitting quality than the distributions with 15 states in most cases.

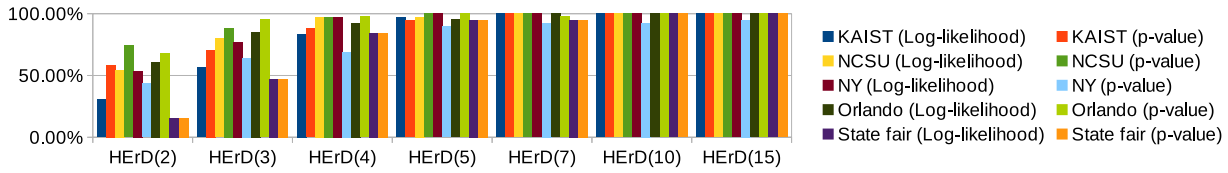


Figure 9: NCSU: Results for distances of users. Percentage of traces where the $HErD(n)$ is better than general distributions according to log-likelihood and p-value.

5 Conclusions

We have presented an extensive comparison for the fitting quality of PHDs and general distribution for various different traces extracted from real world wireless data sets. It was shown, that in most cases even small PHDs with four or five states are sufficient to obtain better results than general distributions widely used in the literature. This could be observed for coarse-grained and fine-grained data. When the trace data exhibits autocorrelation the fitting quality can be further improved by expanding the PHDs into MAPs, although this of course increases the effort for fitting.

For this paper we fitted various characteristics of the traces with distributions. Future work will focus on combining these distributions to obtain realistic mobility and traffic models for wireless networks.

References

[1] S. Asmussen, O. Nerman, and M. Olsson. Fitting phase type distributions via the EM algorithm. *Scand. J. Statist.*, 23:419–441, 1996.

[2] A. Balachandran, G.M. Voelker, P. Bahl, and P. Rangan. Characterizing user behavior and network performance in a public wireless lan. In *Proc. of SIGMETRICS '02*, 2002.

[3] F. Bause, P. Buchholz, and J. Kriege. ProFiDo - The Processes Fitting Toolkit Dortmund. In *Proc. of QEST 2010*, 2010.

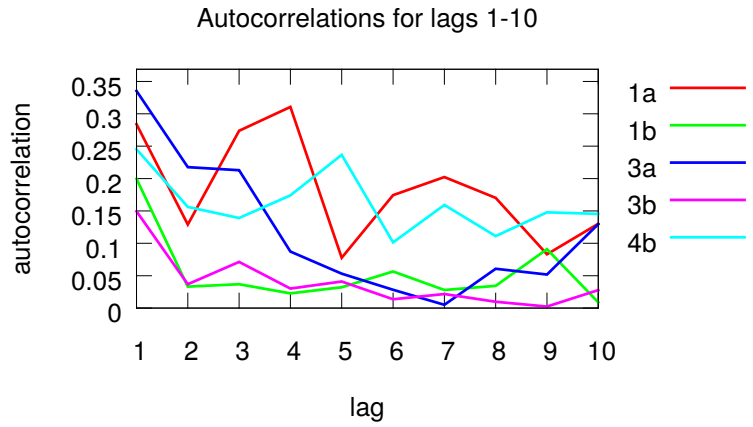
[4] A. Bobbio, A. Horváth, and M. Telek. Matching three moments with minimal acyclic phase type distributions. *Stochastic Models*, 21(2-3), 2005.

[5] P. Buchholz. An EM-Algorithm for MAP Fitting from Real Traffic Data. In *Computer Performance Evaluation / TOOLS*, 2003.

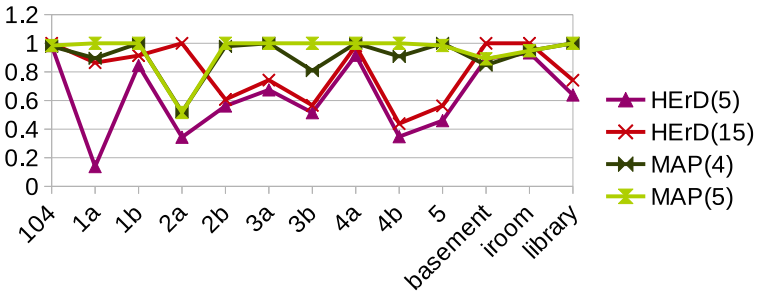
[6] P. Buchholz and J. Kriege. Markov Modeling of Availability and Unavailability Data. In *Proc. of EDCC 2014*, 2014.

[7] P. Buchholz, J. Kriege, and I. Felko. *Input Modeling with Phase-Type Distributions and Markov Models - Theory and Applications*. Springer, 2014.

[8] M. Delignette-Muller and C. Dutang. fitdistrplus: An R package for fitting distributions. *Journal of Statistical Software*, 64(4), 2015.



(a) lag-k autocorrelation



(b) Log-likelihood values for interarrival times at APs (14 traces)

Figure 10: Stanford: Autocorrelation

[9] W. Gao and G. Cao. Fine-grained mobility characterization: Steady and transient state behaviors. In *Proc. of MobiHoc '10*, 2010.

[10] T. Henderson, D. Kotz, and I. Abyzov. The changing usage of a mature campus-wide wireless network. *Computer Networks*, 52(14), 2008.

[11] A. Horváth and M. Telek. Matching more than three moments with acyclic phase type distributions. *Stochastic Models*, 23:167–194, 2007.

[12] G. Horváth. Moment matching-based distribution fitting with generalized hyper-erlang distributions. In *Proc. of ASMTA '13*, 2013.

[13] W. Hsu and A. Helmy. IMPACT: Investigation of mobile-user patterns across university campuses using WLAN trace analysis. Technical report, 2005.

[14] P. Hui, A. Chaintreau, J. Scott, R. Gass, J. Crowcroft, and C. Diot. Pocket switched networks and human mobility in conference environments. In *Proc. of WDTN '05*, 2005.

[15] B. Javadi, D. Kondo, A. Iosup, and D. H. J. Epema. The failure trace archive: Enabling the comparison of failure measurements and models of distributed systems. *J. Parallel Distrib. Comput.*, 73(8):1208–1223, 2013.

[16] M. Kim, D. Kotz, and S. Kim. Extracting a Mobility Model from Real User Traces. In *Proc. of INFOCOM*, 2006.

[17] D. Kondo, B. Javadi, A. Iosup, and D. H. J. Epema. The failure trace archive: Enabling comparative analysis of failures in diverse distributed systems. In *CCGRID*, pages 398–407. IEEE, 2010.

[18] D. Kotz and K. Essien. Analysis of a Campus-Wide Wireless Network. *Wireless Networks*, 11(1), 2005.

[19] D. Kotz and T. Henderson. Crawdad: A community resource for archiving wireless data at dartmouth. *IEEE Pervasive Computing*, 4(4), 2005.

[20] P. L. L'Ecuyer, L. Meliani, and J. Vaucher. Ssj: a framework for stochastic simulation in java. In *Proc. of WSC'02*, 2002.

[21] D. Lelescu, U. Kozat, R. Jain, and M. Balakrishnan. Model t++: An empirical joint space-time registration model. In *Proc. of MobiHoc '06*, 2006.

[22] W. Navidi and T. Camp. Stationary distributions for random waypoint models. *IEEE Transactions on Mobile Computing*, 2004.

[23] M. F. Neuts. A versatile Markovian point process. *Journ. of Appl. Prob.*, 1979.

[24] P. Reinecke, T. Krauß, and K. Wolter. Hyperstar: Phase-type fitting made easy. In *Proc. of QEST'12*, 2012.

[25] I. Rhee, M. Shin, S. Hong, K. Lee, S. J. Kim, and S. Chong. On the levy-walk nature of human mobility. *IEEE/ACM Transactions on Networking*, 19(3), 2011.

[26] D. Tang and M. Baker. Analysis of a local-area wireless network. In *Proc. of MobiCom '00*, 2000.

[27] A. Thümmler, P. Buchholz, and M. Telek. A novel approach for phase-type fitting with the EM algorithm. *IEEE Trans. Dep. Sec. Comput.*, 3(3), 2006.

[28] C. Tuduce and T.R. Gross. A mobility model based on WLAN traces and its validation. In *Proc. of INFOCOM*, 2005.

[29] J. Yoon, B.D. Noble, M. Liu, and M. Kim. Building realistic mobility models from coarse-grained traces. In *Proc. of MobiSys*, 2006.