An Accurate and Analytically Tractable Model for Human Inter-Contact Times

Simon Frohn, Sascha Gübner, and Christoph Lindemann

Department of Computer Science
University of Leipzig
Johannisgasse 26, 04103 Leipzig, Germany

{frohn, guebner, cl}@rvs.informatik.uni-leipzig.de

ABSTRACT

We present an analytically tractable mathematical approach for accurately modeling the distribution of inter-contact times between mobile devices carried by users. The contribution of this paper is two-fold: (1) we show how to employ a Markov-modulated Poisson process (MMPP) for characterizing long-term dependencies in the mobility behavior, and (2) we propose to employ a graph-based clustering approach for taking into account different user groups with inhomogeneous mobility patterns. We illustrate the effectiveness of the proposed approach by considering two comprehensive real-world trace data sets. The presented quantitative results show that the proposed modeling approach closely approximates the dichotomy of the distribution of human inter-contact times into an exponential and power-law distribution observed in recent studies of real-world trace data. As the presented modeling approach for inter-contact times is both analytically tractable and captures long-term dependencies in the mobility behavior, it possesses clear advantages over methods previously introduced for analyzing the performance of opportunistic networking protocols.

Categories and Subject Descriptors

I.6 [SIMULATION AND MODELING];
G.3 [PROBABILITY AND STATISTICS]

General Terms

Measurement, Experimentation

Keywords

Realistic mobility models, mobile ad hoc networks, delay-tolerant networking, Markov-modulated Poisson process

1. INTRODUCTION

With the emergence of mobile communication devices, opportunistic forwarding protocols are becoming an increasingly important area of research. Opportunistic forwarding protocols do not rely on the existence of an end-to-end path to forward a message from a source node to a destination node. Instead, messages are queued and forwarded when devices come into contact. Therefore, the message delay of such protocols strongly depends on the time between two consecutive contact opportunities, i.e. the inter-contact time. To allow quantitative evaluation of such forwarding protocols in early design stages, an accurate analytical model of the inter-contact time distribution is crucial. Existing modeling approaches typically rely on the assumption of exponentially distributed inter-contact times. Also many mobility models, e.g. the random waypoint mobility model [1], deliver exponentially distributed inter-contact times. Recently, large-scale studies of human mobility revealed a dichotomy of the distribution of inter-contact times of mobile users into an exponential and power-law distribution [10]. Therefore, evaluation of opportunistic forwarding protocols based on these mobility models may lead to wrong or incomplete results. To overcome these limitations, several empirical frameworks have been proposed, e.g. [17] that are able to closely fit distributions of inter-contact times derived from traces. The main disadvantage of these models is that they are not analytically tractable. Thus, they have to rely on discrete-event simulation rather than easy to evaluate mathematical models.

In this paper, we present an analytically tractable mathematical approach for accurately modeling the distribution of inter-contact times between mobile devices carried by users. We show how to employ a Markov-modulated Poisson process (MMPP) for characterizing long-term dependencies in the mobility behavior and we propose to employ a graph-based clustering approach for taking into account different user groups with inhomogeneous mobility patterns. To effectively estimate the parameters of the MMPP from measured real-world trace data, we show how to tailor the well-known expectation maximization (EM) algorithm to a numerically stable procedure based on the randomization technique [12]. We illustrate the effectiveness of the proposed approach by considering two comprehensive data sets, i.e., the MIT Bluetooth [6] and UCSD [13] traces. The presented quantitative results show that the proposed modeling approach closely approximates the dichotomy of the distribution of human inter-contact times observed in recent studies of real-world trace data. The presented modeling approach for inter-contact times is both analytically tractable and captures long-term dependencies in the mobility behavior. In particular, the approach could be used to calculate means of delivery times or number of message copies in opportunistic networking [8]. Another application of the presented approach constitutes the generation of synthetic mobility traces, which closely capture the long-term dependencies.

The paper is organized as follows. In section 2, related work is discussed. Section 3 introduces the MMPP approach for modeling inter-contact times, an advanced clustering method and numerically stable parameter estimation. To illustrate the effectiveness of the approach, Section 4 presents quantitative results. Finally, concluding remarks are given.

2. RELATED WORK

Chaintreau et al. [5] showed that the distribution of human inter-contact times exhibits a heavy tail like in a power law distribution. This observation was contrary to widely used constructive mobility models for evaluating ad hoc networking protocols like the random waypoint mobility model. In fact, the construction rules of the random waypoint mobility model and most other widely used mobility models rely on quite some independence assumptions, and, thus, yielding exponentially distributed inter-contact times between nodes. Karagiannis et al. [10] observed that the distribution of inter-contact times first follows a power law and then pass over to an exponential distribution. Similar to [5], we do not assume an exponential distribution for inter-contact times. Opposed to [5] and [10], we propose an analytically tractable approach for modeling the distribution of human inter-contact times.

Cai et al. examined the distribution of inter-contact times for the random walk model in an unbounded area [3]. They showed that for this mobility model the distribution of inter-contact times follows a power law rather than an exponential distribution. The same authors studied in [4] the effect of mobility patterns on the distribution of inter-contact times. They assumed a dichotomy of the inter-contact time distribution, like in [10], and showed that a stronger correlation in the model leads to a more non-exponential head of the distribution. Opposed to [3] and [4], we propose an approach for approximating the distribution of the inter-contact times itself, rather than construction rules for a mobility model.

Srinivasan et al. investigated the contact patterns of students in [16]. They examined the schedules of the students and derived various characteristics. In contrast to this work, we consider contact patterns derived from measured real-world traces instead of contact patterns derived from pre-defined timetables. In [18], Zhang et al. studied the contact patterns between WiFi-equipped busses. They analyzed the aggregated inter-contact times and the inter-contact times on a route-level and derived models to create synthetic traces. Opposed to this work, we focus on the inter-contact times between humans. They aren't constrained to streets and they don't follow fixed routes like the busses in [18].

Yoon et al. proposed a framework in [17], which allows combining wireless access point association traces with an actual map of the considered area. It generates a probabilistic mobility model that produces movement patterns. In a similar manner, Kim et al. proposed an idea in [11] to create an empirical model. It was formed by analyzing information about movement patterns and pause times from a real-world trace measured at Dartmouth College. With the help of this model, they generated synthetic traces and compared them to the real ones. Opposed to [17] and [11], our model, although empirically trained, isn't an empirical model and therefore analytically tractable.

Rhee et al. [15] analyzed GPS traces of persons in various outdoor environments and concluded that many statistical features of human walks can be emulated with a levy walk mobility model. With this model, they recreate inter-contact time distributions observed in other mobility traces. Hsu et al. proposed the time-variant community mobility model TVC [9]. In TVC, a node chooses a community and performs a random direction movement within the bounds of the community. In [14], Mei et al. proposed the mobility model small world in motion. The main idea is that people tend to go to popular places not far away from home. The mobility model was also compared against three real-world traces. In contrast to [9], [14], [15], we focus on the inter-contact time distribution, rather than on a constructive mobility model, as we believe this is the most important metric to conduct performance studies of various protocols. Therefore, it is easier to theoretically analyze it opposed to the complex models proposed in [9], [14] and [15].

3. MODELING HUMAN INTER-CONTACT TIMES USING MARKOV-MODULATED POISSON PROCESSES

To estimate the distribution of inter-contact times between humans, we propose a model based on a Markov-modulated Poisson process (MMPP). We first want to define the inter-contact time that is throughout this paper the aggregated inter-contact time among all fixed node pairs. We refer to this pair wise inter-contact time, as the time between two consecutive starts of a contact among two fixed nodes.

An advantage of the MMPP model is that it is well studied and is analytically tractable. We will further introduce a clustering approach that helps to fit the distribution of the inter-contact times even better and that makes the model more flexible, when used for example in an empirical study of a new protocol.

3.1 Markov-modulated Poisson process

A Markov-modulated Poisson process (MMPP) is a doubly stochastic Poisson process, whose rate varies according to a Markov process [7]. A MMPP is defined through the N-state continuous-time Markov chain (CTMC) with generator matrix $\boldsymbol{Q}$ with

$$\mathbf{Q} = \begin{bmatrix} -\sigma_1 & \sigma_{12} & \cdots & \sigma_{12} \\ \sigma_{21} & -\sigma_2 & \cdots & \sigma_{12} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{N1} & \sigma_{N2} & \cdots & -\sigma_N \end{bmatrix} \quad (1)$$

where $\sigma_i = \sum_{\substack{j=1 \\ j \neq i}}^{N} \sigma_{ij}$ and appropriate N Poisson arrival rates λ_i. To complete the construction of the MMPP, the steady state vector $\boldsymbol{\pi}$ with $\boldsymbol{\pi Q} = 0$ and $\boldsymbol{\pi 1} = 1$, being $\boldsymbol{1}$ the column vector of 1s of appropriate length, is needed. Furthermore, throughout this paper we assume the CTMC to be homogenous, meaning $\boldsymbol{Q}$ and the λ_i are time-independent.

MMPPs have some attractive features that make them suitable to approximate the inter-contact time distribution. In contrast to the exponential distribution, arrivals (or triggered events) derived from a MMPP are not memory-less. The time of the next arrival (meaning the next inter-node contact) depends on the state of the underlying Markov chain. While being in state i for an exponentially distributed time with mean $1/\sigma_i$, the arrivals are exponentially distributed with rate λ_i. The cumulative distribution function of a MMPP is given by

$$F(x) = \boldsymbol{\pi}(\mathbf{I} - e^{(\mathbf{QA})x})(\boldsymbol{\Lambda} - \mathbf{Q})^{-1}\boldsymbol{\Lambda}\mathbf{1} \quad (2)$$

with $\boldsymbol{\pi}$ is the steady state vector, $\boldsymbol{I}$ is the identity matrix, $\boldsymbol{\Lambda}=diag(\lambda_i)$ is a square matrix, and $\boldsymbol{1}=(1,1,\ldots,1)^T$ is a column vector of length N.

3.2 Parameter Estimation of the MMPP

For utilizing the MMPP, we need an estimation algorithm to match the parameters of the generator matrix $\boldsymbol{Q}$ and the Poisson arrival rates λ_i to the characteristics of given trace data. We employ the expectation-maximization (EM) approach, which is a well-known technique to find the appropriate parameters [2], [12]. We used results from previous work [12] tailored to *N*-state MMPPs. In [12], we showed how to employ the EM algorithm for efficiently estimating the parameters for a batch Markovian arrival process (BMAP). The EM algorithm is numerical stable, since the E-step is performed by specially derived computational formulas based on the randomization technique. Since the class of BMAPs includes MMPPs as a special case, we could apply this algorithm by setting the maximum batch size to 1 and restricting the initial parameter set. The parameter estimation based on the EM algorithm and the randomization technique tailored to MMPPs is outlined in the following.

Consider an experiment with an observable part $\boldsymbol{y}$ $\{t_1,...,t_n\}$, i.e. the inter-contact times, and an unobservable part $\boldsymbol{x}$. We want to estimate the parameter set Φ, consisting of the unknown initial state probability vector $\boldsymbol{\pi}$ and the matrices $\boldsymbol{Q}$ and $\boldsymbol{\Lambda}$. $\boldsymbol{D}=\boldsymbol{Q}-\boldsymbol{\Lambda}$ is the rate matrix of transitions without arrivals. $\boldsymbol{\Lambda}$ is the rate matrix of transitions with one arrival. Furthermore, we define the matrix of probability density functions $f(t)$ by

$$f(t) = e^{(Q-\Lambda)t}\Lambda \tag{3}$$

We now can define the likelihood by

$$L(\Phi, \mathbf{y}) = \pi \prod_{k=1}^{n} f(t_k)\mathbf{1} \tag{4}$$

where Φ is the parameter set, $\boldsymbol{y}$ are the observed inter-contact times and $\boldsymbol{1}$ is the column vector of ones of appropriate length.

The EM algorithm iteratively improves the parameter set Φ until a certain criterion is met, like the difference between the consecutive estimates for Φ falls below a threshold or a maximum of iterations is reached. Using the observed information $\boldsymbol{y}$ and the unobserved information $\boldsymbol{x}$, the complete likelihood L^C can be calculated and the expectation step can be conducted numerically stable using Equation (5). This is a modification of Equation (13) in [12]. For the calculation of the other parameters, we refer to [12] due to space limitations. The conditional expectation E_Φ given the estimate Φ is defined by

$$\begin{aligned} E_{\Phi(r)}\{\log(L^c(\Phi,x,y)) \mid N(u), 0 \le u \le T\} &= \sum_{i=1}^{N} \hat{\pi}_i \log \pi_i \\ &+ \sum_{i=1}^{N} D_{i,i}\hat{s}_i + \sum_{i=1}^{N}\sum_{\substack{j=1 \\ j\neq i}}^{N} \hat{T}_{i,j} \log D_{i,j} + \sum_{i=1}^{N} \hat{A}_{i,i} \log \lambda_i \end{aligned} \tag{5}$$

For the special case of a MMPP, $\hat{A}_{i,i}$ reduces to

$$\begin{aligned} \hat{A}_{i,i} = \frac{1}{L(\Phi(r), y)} \sum_{k=1}^{n} \pi(r) \cdot \prod_{l=1}^{k-1} f(t_l) \cdot e^{\mathbf{D}t_k} \\ \cdot \mathbf{1}_i \cdot \lambda_i \cdot \mathbf{1}_i^T \cdot \prod_{l=k+1}^{n} f(t_l) \cdot \mathbf{1} \end{aligned} \tag{6}$$

3.3 Node Clustering

To account for different user groups (e.g. students, staff etc.), we propose to apply a clustering approach for nodes. By dividing the inter-contact times extracted from traces into different clusters, we can model the distribution of inter-contact times among people even better.

The motivation for node clustering is that humans often form groups, e.g. students of a class or colleagues at work. When they do this, they share the same inter-contact patterns, meaning that they meet each other more regular and that they also meet other non-group members similar. Think for example of colleagues, who go together to lunch and meet other persons together.

So our goal is to identify different groups of mobile users resp. nodes in the traces and independently describe their behavior. For this purpose, we first model a graph $G(V,E)$, where $V = \{v_1, v_2, ..., v_n\}$ is the set of n nodes and $E = \{(u,v) \mid u,v \in V, w_{uv} > 0\}$ is the set of weighted links between them. We define T_{uv} as the set of pair wise inter-contact times between two nodes u and v. The link weight w_{uv} is defined as the cardinality of T_{uv}

$$w_{uv} = |T_{uv}| \tag{7}$$

Now we apply a hierarchical clustering approach. Let S_i be the set of j clusters $C_{i,j}$ in the i-th step. We initialize each cluster $C_{0,j}$ with a node

$$C_{0,j} = \{v_j\}, j \in \{1,...,n\} \tag{8}$$

so S_0 is the union of all initial clusters

$$S_0 = \bigcup_{j\in\{1,...,n\}} C_{0,j} \tag{9}$$

Then, in each step, the two clusters with the highest average linking are merged into cluster M

$$\begin{aligned} M = \{C_{i,j} \cup C_{i,k} \mid (C_{i,j}, C_{i,k}) = \\ \underset{\substack{(C_{i,j},C_{i,k})\in S_i \times S_i \\ C_{i,j}\neq C_{i,k}}}{\arg\max} \; avgdist(C_{i,j}, C_{i,k})\} \end{aligned} \tag{10}$$

and the set of clusters for the next step is

$$S_{i+1} = (S_i \setminus \{C_{i,j}, C_{i,k}\}) \cup M \tag{11}$$

Furthermore, we need to define the average linking *avgdist* between two clusters, using the aforementioned link weight w_{uv}, by

$$avgdist(C, C') = \frac{1}{|C||C'|} \sum_{u\in C} \sum_{v\in C'} w_{uv} \tag{12}$$

This procedure continues until only one cluster is left. This type of clustering creates a dendrogram, a graphical representation of the merging steps. For the MIT trace, we could identify four different clusters. We visualized the results in Figure 1 by coloring the nodes of a cluster. In the graph, each node is connected with its 3 most seen neighboring nodes.

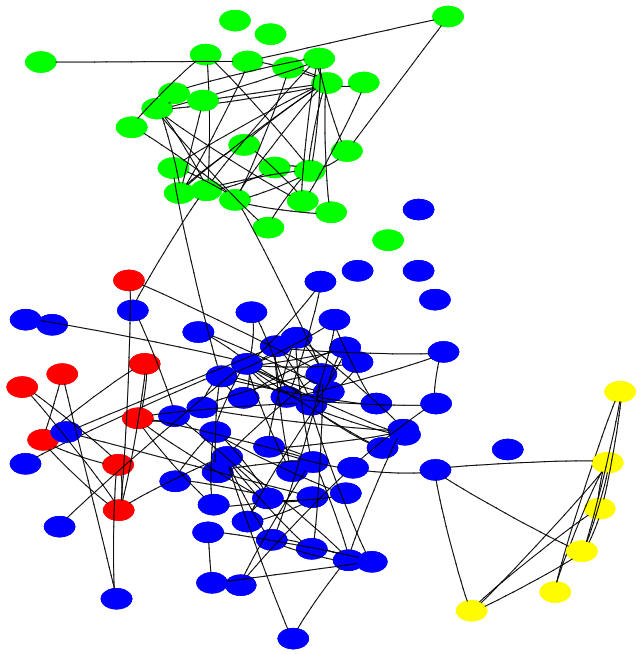

Figure 1. Illustration of the node clustering approach for the MIT trace [6]

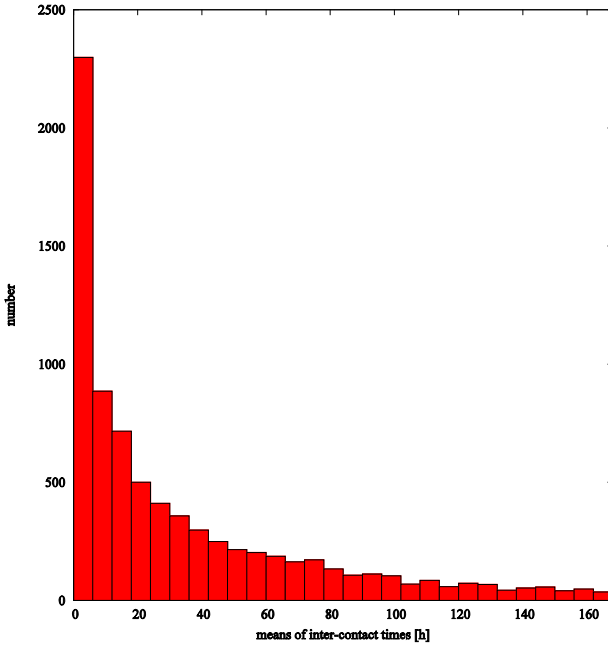

Figure 2. Histogram of the means of the inter-contact times for UCSD trace [13]

The resulting node clusters match very well with clusters found in previous work [6]. The clusters are used to divide the inter-contact times of the trace into sets. We achieve this by collecting all pair wise inter-contact times for a given node. Therefore, we define the set of inter-contact times $\tilde{T}_j$ for a cluster C_j as

$$\tilde{T}_j = \bigcup_{\substack{u \in C_j \\ v \in V}} T_{uv} \tag{13}$$

To further optimize our clustering approach, we also cluster node pairs dependent on their mean value of their pair wise inter-contact times. The distribution of these mean values can be seen in Figure 2 for the UCSD trace set. We define the mean inter-contact time m_{uv} of a node pair by

$$m_{uv} = \frac{\sum_{t \in T_{uv}} t}{w_{uv}}, u, v \in V \tag{14}$$

Subsequently, according to this mean value we group the node pairs into different classes. We choose these classes to reflect important time distances in the periodic behavior of human interacting, like one hour, one day or one week. Then, we model the inter-contact times for each class with an MMPP whose parameters are estimated by Eq. (5) and (6).

An algorithmic description of the presented node clustering approach is given in Figure 3 where the initialization step uses Eq. (8) and (9), and the clustering step uses Eq. (10), (11) and the function *avgdist* defined in (12). We further use a distance matrix ***D*** to hold the computed average linking distances between all clusters. In each step two clusters are merged and the matrix ***D*** is decreased in size and entries concerning the two merged clusters are recalculated. This saves computation time through avoiding unnecessary calculations.

```
1:   Initialization step
2:   S0 = ∅
3:   foreach node vi ∈ V do
4:       Ci = {vi}
5:       S0 = S0 ∪ Ci
6:   done
7:   foreach cluster C ∈ S0 do
8:       foreach cluster C'∈ (S0 \ C) do
9:           D(C,C') = avgdist(C,C')
10:      done
11:  done
12:
13:  Clustering step
14:  for 0≤i<n do
15:      cluster_pair_max = (∅,∅)
16:      cluster_dist_max = 0
17:      foreach cluster C ∈ Si do
18:          foreach cluster C'∈ (Si \ C) do
19:              if (D(C,C') > cluster_dist_max)
20:                  cluster_pair_max = (C,C')
21:                  cluster_dist_max = D(C,C')
22:              endif
23:          done
24:      done
25:      M = cluster_pair_max,1 ∪ cluster_pair_max,2
26:      Si+1 = (Si \ { cluster_pair_max,1 , cluster_pair_max,2}) ∪ M
27:      Delete row/column cluster_pair_max{1,2} in D
28:      Add row/column M to D
29:      foreach cluster C ∈ (Si+1 \ M) do
30:          D(M,C) = D(C,M) = avgdist(M,C)
31:      done
32:  done
```

Figure 3. Algorithmic description of the clustering approach

4. Evaluation

To evaluate the proposed MMPP model, we compare the inter-contact times derived from our model to the inter-contact times of two comprehensive real-world traces, namely the MIT Bluetooth trace [6] and traces collected through the WTD project at UCSD [13]. Table 1 gives a short summary of some key values of the data. The MIT trace covers over 114,046 contacts between 96 devices collected over 283 days. The UCSD trace lasts 77 days, covering over 268,899 contacts among 250 devices.

We show that our model fits the distribution of the inter-contact times quite well and that this is even improved through the use of clustering. Additionally, we compare our model with an exponential and a power-law distribution and show that these distributions approximate the distribution of human inter-contact times very inaccurately.

4.1 Trace Data

Both data sets have a huge number of participating nodes, a long running time, but also a fine grained scan interval. This is important, because we wanted a representative amount of data where devices are worn by people to derive the inter-contact times between humans.

In both datasets different types of contact tracing can be found. For the MIT experiment, users carried mobile phones (Nokia 6600) where the Bluetooth devices scan their proximity (approximately 5-10m) periodically and log seen Bluetooth IDs. The mobile phones were given to students or faculty members of MIT Media Laboratory. 25 of them were given to incoming students of MIT Sloan business school.

We had to map the device Ids with the person Ids, because some persons wore more than one device. Furthermore, note that contacts weren't always bidirectional, as some contacts were only seen by one partner. We also had to omit one device for our studies, as it hadn't seen any other devices and also wasn't seen by any others.

For the UCSD trace the derivation of the inter-contact times is conducted differently. In this study WiFi enabled PDAs were given to 275 freshmen at the campus. A preinstalled software scans periodically for access points (not only the connected one) and logs their Ids. We derived the inter-contact times by assuming that two devices can communicate with each other, when they share one or more access points. Although this is perhaps too optimistic, these assumptions were also made in previous work, like [10].

Finally, we want to remark that both data sets share some common types of error sources that might be taken into account. These errors include:

- devices may not always been carried by the user
- devices may not always been activated
- devices may miss some contacts (because of the used scan interval or other outer influences)

We think that these error types don't influence our study significantly and can be neglected. The CCDF in log-log scale of the derived inter-contact times from both data sets can be seen in Figure 4.

Table 1. Considered Traces

	MIT-Trace	UCSD-Trace
Network type	Bluetooth	WiFi
Devices	96	250
Duration	283 days	77 days
Contacts	114,046	268,899
Scan interval	300 sec	20 sec
Mean of inter-contact times	85 h	19 h

4.2 Methodology

We extracted the pair wise inter-contact times from both traces first. As defined in section 3, the pair wise inter-contact times are the intervals between consecutive beginnings of a meeting between two fixed nodes. While we could extract this information almost directly out of the MIT traces, we had to transform the UCSD trace to get the appropriate measurements. Therefore, we defined sessions where clients "see" an access point, as described in [13]. Note that the clients log not only access points with which they are actually associated, but also access points from which they receive beacons. We used the sessions to characterize time durations where two clients see each other, in particular, when they share an access point. Because clients could see each other through different access points over time, we needed to combine these results to model client inter-contact sessions in the same manner as for the MIT trace set.

We aggregated these pair wise inter-contact times and computed the empirical cumulative distribution function (CDF) resp. the empirical complementary cumulative distribution function (CCDF) for both traces. Then we utilize the clustering approach on the data.

To train our model on the clustered and unclustered data we considered a 2-state MMPP, denoted as MMPP-2, as they are already capable to model the conducted distributions very well. In case of a MMPP-2 $\boldsymbol{Q}$ and $\boldsymbol{\pi}$ are given by

$$\mathbf{Q} = \begin{bmatrix} -\sigma_1 & \sigma_1 \\ \sigma_2 & -\sigma_2 \end{bmatrix} \tag{15}$$

$$\boldsymbol{\pi} = (\pi_1, \pi_2) = \frac{1}{\sigma_1 + \sigma_2}(\sigma_2, \sigma_1) \tag{16}$$

The results of our parameter estimation approach can be seen in Table 2.

Table 2. Estimated MMPP parameters for both traces

	MIT-Trace	UCSD-Trace
σ_1	1.0465	0.1430
σ_2	0.0949	0.0055
λ_1	2.9717	1.2301
λ_2	0.0412	0.0085

Table 3. Deviation of the different approaches acc. to (18)

	MIT-Trace	UCSD-Trace
Exponential distribution	17.86	66.18
Power-law distribution	0.7259	0.6181
MMPP w/o clustering	0.8189	0.4508
MMPP w/ clustering	0.4333	0.3686

Table 4. Deviation of the different approaches acc. to (19)

	MIT-Trace	UCSD-Trace
Exponential distribution	0.4180	0.6139
Power-law distribution	0.1482	0.1232
MMPP w/o clustering	0.1497	0.0629
MMPP w/ clustering	0.0891	0.0392

In order to get inter-contact times distributed according to a MMPP-2, we evaluate the underlying Markov chain. We create two independent streams of points of time, whose distance is exponentially distributed with rate λ_i. We slice through the streams by simulating the underlying CTMC defined through $\boldsymbol{Q}$. As long as the CTMC is in state i, the arrivals occur according to stream i. From the so formed sequence of times, we derive the inter-contact times by simply subtracting consecutive times.

To visualize the results of this process, we created synthetic data out of the different models. To take the different cluster sizes into account, we combined the data produced for each cluster in a weighted manner. Then the empirical CDF resp. CCDF was computed for all traces.

We also compare our results with an exponential and a power-law distribution. To fit the exponential distribution to the empirical data, we set the rate λ to $1/m$ where m is the mean value of the aggregated trace inter-contact times. For the power-law distribution [10], we used linear regression in the CCDF to set t_0 and α accordingly.

We used two methods to measure the deviation from trace data. First we used a Chi-Squared based method. We sorted the original data traces into n equally sized (in terms of number of inter-contact times from the original trace) bins $B^T{}_i$. We then sorted the synthetic data of the unclustered model and that of the clustering approach into these bins, denoting their entries as $B^S{}_i$. Furthermore, let m^T be the number of all inter-contact times of the original trace and m^S be the number of all inter-contact times of the considered synthetic trace. We further define $P(B^X{}_i)$

$$P(B_i^X) = \frac{\left|B_i^X\right|}{m^X}, X \in \{S,T\}, 1 \le i \le n \qquad (17)$$

as the probability for a random chosen value to be in bin i in the original trace resp. in the synthetic trace. We sum up the squared differences of the probability for a value being in bin B_i divided by the expected probability observed in the original trace for each class. The failure F_{X^2} is then defined by

$$F_{X^2} = \sum_{i=1}^{n} \frac{(P(B_i^S) - P(B_i^T))^2}{P(B_i^T)} \qquad (18)$$

Furthermore, we calculated the maximum difference between the empirical CDFs based on the Kolmogorov-Smirnov (KS) test. Let M^T resp. M^S be the set of all inter-contact times in the original resp. synthetic trace. Then the failure F_{KS} is defined by

$$F_{KS} = \max_{x \in M^T \cup M^S} \left|F^T(x) - F^S(x)\right| \qquad (19)$$

with F^T resp. F^S being the empirical CDF of the original resp. synthetic trace data.

The results of the Chi-Square based method can be seen in Table 3. We set the number of bins n to 50. We can observe that the exponential distribution doesn't fit the distributions of the inter-contact times very well for both traces. The power-law distribution fits quite like the MMPP model, but both are outperformed by the clustering approach.

A similar observation for the Kolmogorov-Smirnov based method can be made in Table 4 where the maximum differences in the distributions are depicted. We see that the exponential distribution is again inaccurate. The power-law distribution is in the same order as the MMPP model. Again, the clustering approach fits best.

4.3 Quantitative Results

We first examine the results of our investigation of the MIT trace. As it can be seen in Figure 6 (lin-lin scale), half of the contact intervals are smaller than 12 hours. As we can clearly see the exponential distribution heavily underestimates this part of the inter-contact time distribution. The power-law distribution fits quite well in the first 12 hours. This is consistent with the findings in [10]. It can also be seen that our clustering approach fits very good. In Figure 8 we see the overall distribution of the inter-contact times in log-log scale. There we see the main disadvantage of the power-law distribution. It decreases linearly, leading to a heavy tail of the distribution, which is the reason why it is hardly analytically tractable. The expected value of such a distribution is infinite. Contrary to this, the MMPP model and the clustering approach still fit the distribution when reaching the tail.

Additionally, Figure 5 depicts the distributions of the found clusters for the MIT trace determined by the clustering algorithm. The figure shows that the node clusters indeed behave different with respect to the distribution of the inter-contact times.

When we examine the inter-contact times of the UCSD data in Figure 7 (lin-lin scale) and Figure 9 (log-log scale), we can see a similar observation. Again, the exponential distribution model is a poor approximation for the distribution of the inter-contact times. It underestimates the first 12 hours and falls down too fast in the tail of the distribution. Although a good estimation for the first 12 hours, the power-law distribution can't be used after that time span. Again, the MMPP model serves as a very good approximation for the distribution, although it falls down a little too fast in the tail. The clustering method is the best approach. It estimates both very well, the head, i.e. the first 12 hours, and the tail of the distribution.

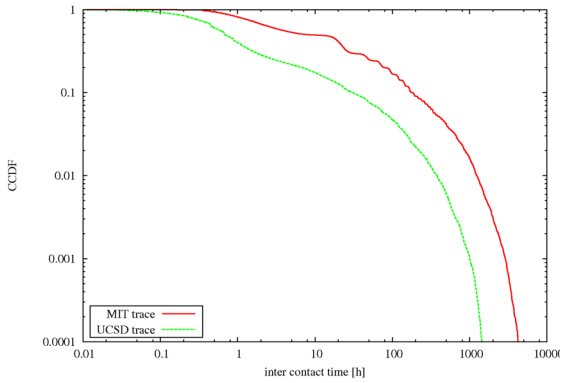

Figure 4. CCDF of inter-contact times for both traces.

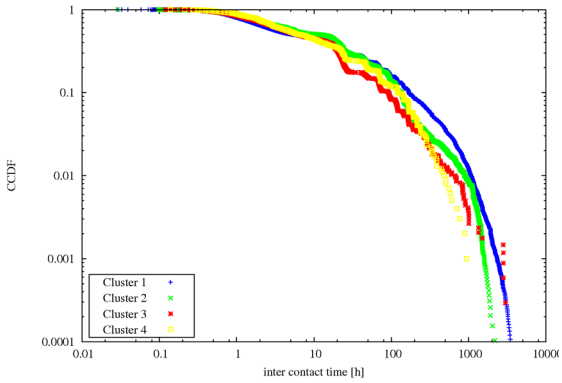

Figure 5. CCDF of inter-contact times for clusters found in MIT trace.

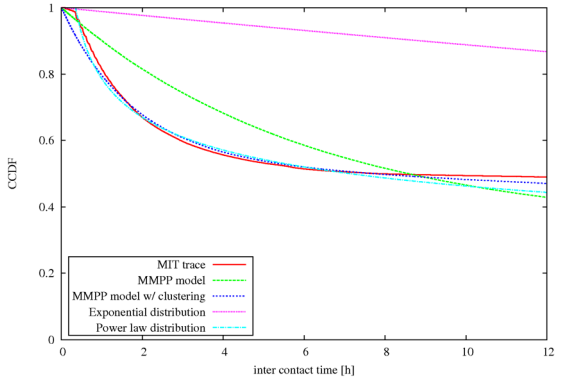

Figure 6. CCDF of inter-contact times for MIT trace (first 12h).

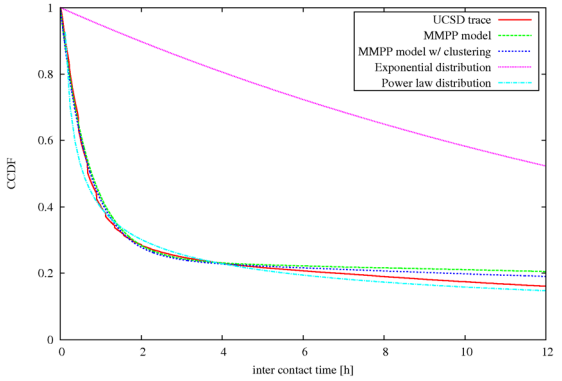

Figure 7. CCDF of inter-contact times for UCSD trace (first 12h).

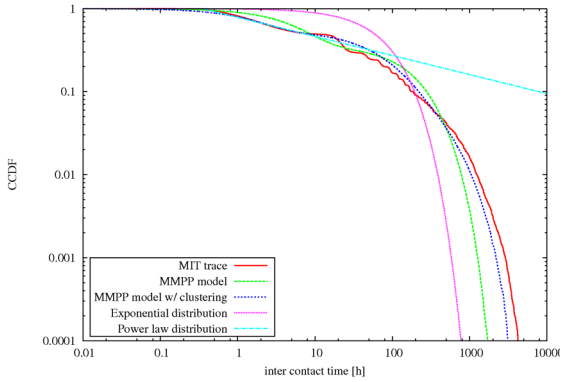

Figure 8. CCDF of inter-contact times for MIT traces.

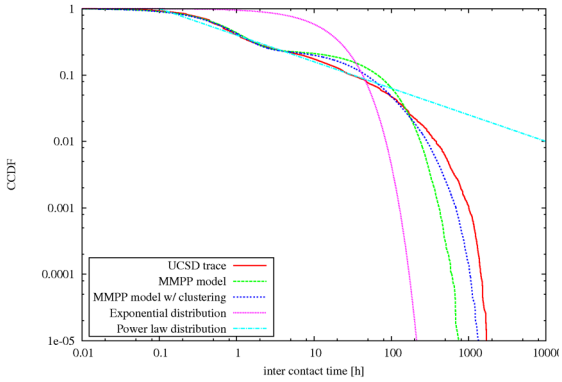

Figure 9. CCDF of inter-contact times for UCSD traces.

5. CONCLUSION

In this paper, we presented an analytically tractable mathematical approach for modeling the distribution of human inter-contact times. We showed how a Markov-modulated Poisson process can be employed to characterize the long-term dependencies in the patterns of human contacts. We further proposed a graph-based clustering approach that takes different user groups with inhomogeneous mobility patterns into account. By comparing our model with two comprehensive real-world trace data sets, we further showed that our approach is a good choice to estimate the distribution of human inter-contact times and that it is able to closely approximate the dichotomy of this distribution, while still being analytically tractable.

For future work we plan to employ our MMPP model to quantitatively evaluate new opportunistic forwarding protocols.

6. REFERENCES

[1] Bettstetter, C., Hartenstein, H., and Pérez-Costa, X. 2004. Stochastic properties of the random waypoint mobility model. *Wirel. Netw.* 10, 5 (Sep. 2004), 555-567.

[2] Buchholz, P. 2003. An EM-Algorithm for MAP Fitting from Real Traffic Data. *Proc. Int. Conf. On Modelling Techniques and Tools for Computer Systems Performance Evaluation*, LNCS Vol. 2794 (Sept. 2003), 218-236

[3] Cai, H. and Eun, D. Y. 2007. Crossing over the bounded domain: from exponential to power-law inter-meeting time in MANET. In *Proceedings of the 13th Annual ACM international Conference on Mobile Computing and Networking* (Montréal, Québec, Canada, September 09 - 14, 2007). MobiCom '07. ACM, New York, NY, 159-170.

[4] Cai, H. and Eun, D. Y. 2008. Toward stochastic anatomy of inter-meeting time distribution under general mobility models. In *Proceedings of the 9th ACM international Symposium on Mobile Ad Hoc Networking and Computing* (Hong Kong, Hong Kong, China, May 26 - 30, 2008). MobiHoc '08. ACM, New York, NY, 273-282.

[5] Chaintreau, A., Hui, P., Crowcroft, J., Diot, C., Gass, R., and Scott, J. 2007. Impact of Human Mobility on Opportunistic Forwarding Algorithms. *IEEE Transactions on Mobile Computing* 6, 6 (Jun. 2007), 606-620.

[6] Eagle, N., and Pentland, A., CRAWDAD data set mit/reality (v. 2005-07-01), July 2005

[7] Fischer, W. and Meier-Hellstern, K. 1993. The Markov-modulated Poisson process (MMPP) cookbook. *Perform. Eval.* 18, 2 (Sep. 1993), 149-171

[8] Groenevelt, R., Nain, P., and Koole, G. 2005. The message delay in mobile ad hoc networks. *Perform. Eval.* 62, 1-4 (Oct. 2005), 210-228.

[9] Hsu, W., Spyropoulos, T., Psounis, K., and Helmy, A. 2009. Modeling spatial and temporal dependencies of user mobility in wireless mobile networks. *IEEE/ACM Trans. Netw.* 17, 5 (Oct. 2009), 1564-1577.

[10] Karagiannis, T., Le Boudec, J., and Vojnović, M. 2007. Power law and exponential decay of inter contact times between mobile devices. In *Proceedings of the 13th Annual ACM international Conference on Mobile Computing and Networking* (Montréal, Québec, Canada, September 09 - 14, 2007). MobiCom '07. ACM, New York, NY, 183-194.

[11] Kim, M., Kotz, D., and Kim, S. 2006. Extracting a mobility model from real user traces, in *Proceedings of the 25th Annual Joint Conference of the IEEE Computer and Communications Societies.* (Barcelona, Spain, April 2006). INFOCOM '06.

[12] Klemm, A., Lindemann, C., and Lohmann, M. 2003. Modeling IP traffic using the batch Markovian arrival process. *Perform. Eval.* 54, 2 (Oct. 2003), 149-173.

[13] McNett, M. and Voelker, G. M. 2005. Access and mobility of wireless PDA users. SIGMOBILE Mob. Comput. Commun. Rev. 9, 2 (Apr. 2005), 40-55.

[14] Mei, A., and Stefa, J. 2009. Swim: A simple model to generate small mobile worlds. In *Proceedings of the 28th Conference on Computer Communications* (April 2009). IEEE INFOCOM '09, 2106-2113

[15] Rhee, I., Shin, M., Hong, S., Lee, K., and Chong, S. 2008, On the Levy-Walk nature of human mobility, in *IEEE Conference on Computer Communications.* (April 2008). IEEE INFOCOM `08., 924-932.

[16] Srinivasan, V., Motani, M., and Ooi, W. T. 2006. Analysis and implications of student contact patterns derived from campus schedules. In *Proceedings of the 12th Annual international Conference on Mobile Computing and Networking* (Los Angeles, CA, USA, September 23 - 29, 2006). MobiCom '06. ACM, New York, NY, 86-97.

[17] Yoon, J., Noble, B. D., Liu, M., and Kim, M. 2006. Building realistic mobility models from coarse-grained traces. In *Proceedings of the 4th international Conference on Mobile Systems, Applications and Services* (Uppsala, Sweden, June 19 - 22, 2006). MobiSys '06. ACM, New York, NY, 177-190.

[18] Zhang, X., Kurose, J., Levine, B. N., Towsley, D., and Zhang, H. 2007. Study of a bus-based disruption-tolerant network: mobility modeling and impact on routing. In *Proceedings of the 13th Annual ACM international Conference on Mobile Computing and Networking* (Montréal, Québec, Canada, September 09 - 14, 2007). MobiCom '07. ACM, New York, NY, 195-206.