Modeling of a Bus-based Disruption Tolerant Network Trace

Xiaolan Zhang, Jim Kurose, Brian Levine, Don Towsley, Honggang Zhang
University of Massachusetts [§]Suffolk University
{ellenz,kurose,brian,towsley@cs.umass.edu} hzhang@suffolk.edu

August 13, 2007

UMass Computer Science Technical Report 2007-12

Abstract

We study traces taken from UMass DieselNet, a Disruption-Tolerant Network consisting of WiFi nodes attached to buses. As buses travel their routes, they encounter other buses and in some cases are able to establish pair-wise connections and transfer data between them. We analyze the bus-to-bus contact traces to characterize the contact processes between buses and its impact on DTN routing performance.

We find that the all-bus-pairs aggregated inter-contact times show no discernible pattern. However, the inter-contact times aggregated at a route level exhibit periodic behavior. Based on analysis of the deterministic inter-meeting times for bus pairs running on route pairs, and consideration of the variability in bus movement and the random failures to establish connections, we construct generative route-level models that capture the above behavior, allowing one to generate synthetic DTN mobility traces. Through trace-driven simulations of epidemic routing, we find that the epidemic performance predicted by traces generated with this finer-grained route-level model are much closer to the actual performance that would be realized in the operational system than traces generated using the coarse-grained all-bus-pairs aggregated model. This suggests the importance in choosing the right level of model granularity when modeling mobility-related measures such as inter-contact times in DTNs.

1 Introduction

The many advantages offered by mobile communications have pushed wireless networks beyond supporting laptops in buildings to more challenging environments. Many of the wireless mobile ad hoc networks for groups of vehicles, pedestrians, or tracked wildlife experience intermittent node connectivity and disconnection of nodes or groups of nodes due to limitations of power, mobility, node density, and equipment failure. Network architecture and protocol designs that route data despite intermittent connectivity among nodes are generally referred to as Disruption-Tolerant Networks (DTNs). Such networks have been deployed in the context of buses [6, 4, 28], pedestrians [7], animal tracking [19], and underwater sensor networks [10].

Unlike other network regimes — such as tethered networks or multi-hop, unpartitioned MANETs — routing performance in DTNs is primarily affected by the frequency and duration of opportunities for data transfer between nodes. Therefore, when studying the performance of routing protocols and applications in DTNs, it is important to have models that accurately characterize these transfer opportunities.

There is a rich body of work on the measurement, characterization, and modeling of mobility traces taken from contemporaneously connected wireless LANs [31, 17, 22] and mobile ad hoc networks [18, 3]. Several recent studies have characterized traces collected from actual mobile networks with intermittent connectivity [7, 6] or adapted from traces collected from wireless LAN, and evaluate the impact of the measured mobility on DTN applications [29, 16]. These works characterized only certain aspects of traces, e.g., the *aggregate* inter-contact time or the contact graph, without considering which aspects of the underlying mobility patterns are most important in determining DTN performance and therefore need to be captured and modeled accurately.

In this work, we develop a *generative* model of the inter-contact time of DTN nodes based on traces collected from our operational vehicular DTN, UMass DieselNet [6]. The model is generative in that it can be used to generate synthetic traces of node inter-contact times that can then be used to drive simulations. As we will see, however, these models are of interest in their own right, as models at the appropriate level of granularity can reveal structure that is hidden at the aggregate level and that can influence DTN performance. Indeed, a focus of our research is to understand the right level of modeling granularity so that traces generated by the model can then be used in simulation to accurately predict DTN performance. We show that while the all-bus-pairs aggregated inter-contact times show no clear pattern, inter-contact times at the bus-route level show periodic structure that can be modeled as mixtures of normal distributions (whose parameters can be inferred from empirical traces using an EM algorithm). Using a trace-driven simulation of epidemic routing, we show that this finer-grained route-level model of inter-contact times predicts performance much more accurately than the coarser-grained aggregated all-bus-pairs model.

The remainder of this report is structured as follows. Section 2 describes our testbed and trace data. In Section 3, we describe the performance metrics that we use to evaluate our generative model. In Section 4, we evaluate the aggregate model used in previous work and show that it does not perform well by our metrics. In Section 5 we propose a route-level model that generates synthetic traces that better match the routing performance of the original trace. We review related works in Section 6, and summarize this technical report and discuss future work in Section 7.

2 UMass DieselNet Traces

In this section, we first describe the UMass DieselNet [6] testbed, explain the traces collected and present background information about bus dispatching. We then describe our preprocessing of the traces, including merging two directional processes into one symmetric process, merging contacts that occurs close in time, and excluding falty devices from the traces.

2.1 Testbed and Trace Collection

UMass DieselNet consists of 40 buses serving the area surrounding the University of Massachusetts, Amherst campus. Each bus is equipped with a Linux computer, an 802.11b Access Point (AP), a second 802.11b interface, and a GPS device. The AP on each bus beacons its SSID once every 100 ms. The second radio continuously searches for SSID broadcasts. On discovering a remote bus's AP, the discovering bus obtains an IP address from the remote bus. Then, a TCP connection is opened, initiating a *contact event*, and data is continuously transmitted to the remote bus until the TCP connection is broken when the buses move out of range. Once the socket reports an error or closure, the contact event is marked as ended and logged. For each contact, the receiver logs the ID of the sender, the time, duration, and the number of bytes received. These *bus-to-bus transfer records* are transmitted to a central repository whenever a bus is able to associate with a fixed 802.11 access point that is attached to the Internet (e.g., offered by a cafe or in the bus garage). We refer to the records of the times and locations that each bus connects to fixed APs as *(bus-to-AP) check-in records*.

It is helpful to understand how buses are scheduled and dispatched since these are the primary determinants of bus mobility. The bus system serves approximately ten *routes*. Some routes have more buses running at the same time than others. In this work, we focus on the three most popular routes, the campus SHUTTLE that tours the campus in a butterfly shape route (see Section 5.2 for details) and the SN_SA and NA_BR routes that travel between our campus and nearby towns within 150 square miles. During weekdays, beginning at approximately 7 A.M. and ending at approximately 7 P.M., there are multiple *shifts* serving each of these three routes. The shifts within a route are spaced so that there is a 15-minute spacing between shifts. Each of the other seven routes is served by only one or two shifts at a time, or they may be served only every other day, and in general we have fewer data points characterizing their operation.

For dispatching and driver assignment purposes, shifts are divided into morning (AM), midday (MID), afternoon (PM), and evening (EVE) *sub-shifts*. In the morning, drivers choose buses at random to run the AM sub-shifts. At the end of the AM sub-shift, the bus is often handed over to another driver (often at a bus stop) to operate the next sub-shift; but in some cases, the bus returns to the bus garage, and it is then possibly assigned to another shift on that route or to another route.

Our results are based on the study of 55 days of traces collected during the spring 2006 semester, from Jan 30 to May 28 with weekends, spring break, and holidays removed since during these times the buses run on reduced schedules. We focus on the events logged between 7 A.M. and 7 P.M. for each day, when buses are running regularly. We also use bus dispatching records, which record the mapping from buses to routes and shifts for each day. Both the traces and dispatching records are available for download at http://traces.cs.umass.edu.

2.2 Mobility Traces Preprocessing

As discussed earlier, when two buses are in transmission range, each one connects to the other's AP to transmit data to the other bus using a separate TCP connection, i.e., the recorded contacts are directional. The contacts in both directions are over the same 802.11b channel; as a result, during one physical meeting of two buses, there can be multiple directional contacts (in both directions) as they gain or lose access to the shared channel. We note that this is due to the way the current system is built; we can imagine systems where symmetric contact is established when two buses meet or where two different channels are used for the contacts in each direction. As we wish to focus on the mobility rather than the specific operations of the bus nodes (including MAC layer), we assume all contacts are symmetric, i.e., data can flow in either direction.

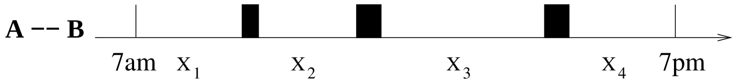


Figure 1: A contact process between bus A and bus B. Here X_2, X_3 are fully observed inter-contact time, X_1 is a start-censored observation, and X_4 is a end-censored observation of inter-contact time

	transfer & check-in	transfer & no check-in	no transfer & check-in	no transfer & no check-in
Counts	1055	42	34	228
%	77.91	3.1	2.5	16.84

Table 1: Number and fraction of four different cases of daily bus status. There are 1,354 records in total.

Figure 1 illustrates the contact process between two buses, A and B, during a day. In the figure, we use black boxes to represent contacts and spaces in-between to represent the interval between contacts. We refer to the duration of time between two subsequent contacts as the *inter-contact time*.

In our traces, there are many very short inter-contact times. This occurs, for example, when two buses that travel closely in the same direction repeatedly come in and out of range of each other as their spacing changes with the road traffic. We merge such events as little else can occur between the contacts. Specifically, for each pair of buses, we combine any two subsequent contacts that occur within 60 seconds of each other. The merged contact has a starting time equal to the earlier contact's starting time and an ending time equal to the later contact's ending time.

We observed that some buses operating on routes during a day were not observed in the traces. It may be the case that the bus did not physically meet other buses, but it may also be the case that the bus failed to set up TCP connections when in range of other buses. There are several reasons for the latter. When the buses are moving at high speeds, there is not enough time for two passing buses to form an 802.11 association and initiate a TCP transfer; we have data on bus speeds that confirms this problem. Hardware failures are not uncommon on the testbed and occasionally mechanics disable the system when servicing the bus and neglect to enable it afterwards. As we don't know whether devices were functioning correctly when the traces were collected, for any day, if we observed a bus in the bus-to-bus transfers or bus-to-AP check-in records, then we assume the device on the bus worked properly for that whole day; otherwise, the device was assumed to be faulty and the bus was removed from the trace for the day.

Table 1 shows, among all the buses running on routes during the whole trace, the numbers and fractions of instances that a bus *(i)* had transfer records and check-in records, *(ii)* had transfer records but no check-in records, *(iii)* had no transfer records but had check-in records, and *(iv)* had no transfer records or check-in records during a day. We observe that the correlation of "having a bus-to-bus contact" and "having check-in records" is high.

3 Performance Characteristic under the Trace

The goal of modeling the bus mobility trace is to correctly predict DTN routing performance. In the past, many routing schemes have been proposed for DTNs, but we focus here on basic epidemic routing. When there are no resource constraints in the network, epidemic routing provides the best-case delivery delay performance. In this section, we first introduce epidemic routing and the performance metrics that we are interested in, and then describe the trace-driven simulation we develop for evaluating the performance of epidemic routing under given mobility traces.

3.1 Epidemic routing and performance metrics

Epidemic routing [32] adopts a "store-carry-forward" paradigm: a node receiving a packet buffers and carries that packet as it moves, passing the packet on to new nodes that it encounters. Analogous to the spread of infectious diseases, each time a packet-carrying node encounters a new node that does not have a copy of that packet, the carrier is said to *infect* this new node by passing on a packet copy; newly infected nodes, in turn, behave similarly. The destination receives the packet when it first meets an infected node, and initiates a recovery process that delete packets copies at infected nodes by the propagation of acknowledgment information in the network. Many recovery schemes have been previously proposed and studied [33, 6]; we will adopt the VACCINE recovery scheme in which acknowledgment information is propagated maximally in the same manner as data packet.

For DTN routings, there exists a trade-off between packet delivery delay and the number of copies made for each packet, where delivery delay is an important performance metric for application and the number of copies made is directly related to transmission overhead. In this work, in addition to these two important performance metrics, we also study the *number of hops of the minimal delay paths discovered by epidemic routing*. This hop count metric is useful in setting the maximum number of hops in a K-hop scheme [12].

3.2 Trace-driven Simulation of Epidemic Routing

As our primary focus is on the impact of mobility on DTN routing, we assume there is no resource contention in the network in terms of bandwidth or buffers and that when two buses come into contact, they can instantaneously exchange an arbitrary number of packets. We next describe the trace-driven simulation that we use evaluate the performance of epidemic routing under a given mobility trace under these assumptions.

A meeting trace can be represented as $G =< V, L >$, where V is the set of nodes and L is the set of edges. Each edge $l \in L$, represents a contact between two nodes $v_1, v_2 \in V$, and is labeled with the time interval that the contact happens, $[s(l), e(l)]$, where $s(l)$ is the starting time of the contact, and $e(l)$ is the ending time of the contact.

Under instantaneous transmission assumption, as Figure 2 demonstrates, in order for a packet generated at node *src* at time t to reach the destination node *dest*, a time-respective path in the network, $P = (l_1, ..., l_k)$, is required such that $e(l_1) \geq t$. A path is called time-respective path if the edges along the path have increasing time labels, i.e., $s(l_i) < e(l_j)$, for any $i < j$. Each path is available until a certain time, i.e., $avail(P) = min\{e(l_i), i = 1, ..., k\}$. A packet generated at time t traversing along path P will experience a delay given by $max\{0, s(l_i) - t, i = 1, ..., k\}$. This means that as t increases, the delay on a path decreases linearly with time, until it becomes 0 after which the delay remains 0 until $t = avail(P)$. We assume that packets generated at $t = avail(P)$ can be delivered through path P, but packets generated *right after* this time, denoted as $avail(P)+$, cannot be delivered through path P.

We wish to evaluate the delivery delay, the number of copies made, and the hop count of the path for packets generated *at any point of time for given source-destination pair* under epidemic routing scheme. We next consider the plots that depicts such information, e.g., the packet-delivery-delay versus packet-generation-time plot and the number-of-copies-made versus packet-generation-time plot. The following observations allow us to simulate the propagation of a finite number of packets to obtain these information.

First, one can show that, for a class of routing schemes including epidemic routing, the delivery-delay

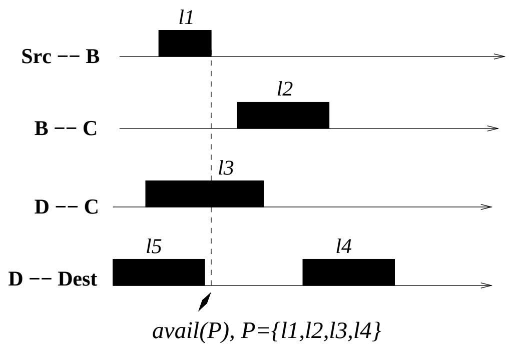
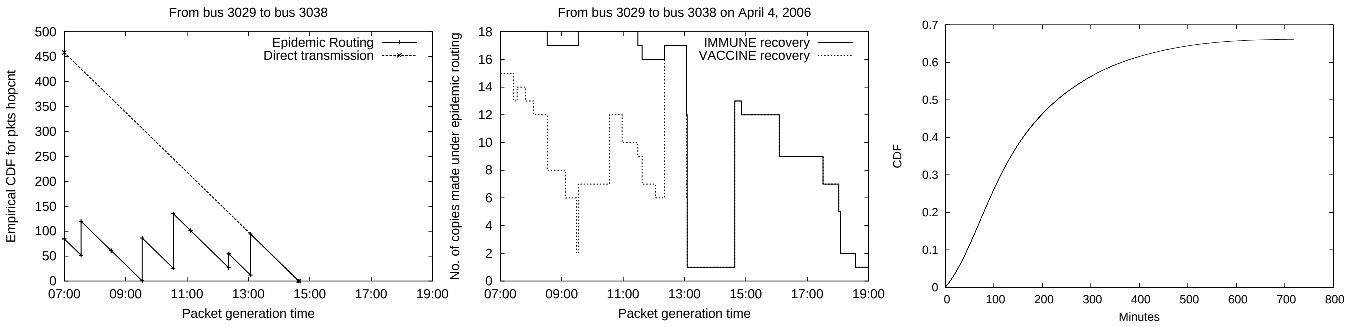


Figure 2: Example of a path $P = \{l_1, l_2, l_3, l_4\}$ from src to dest. Note that $\{l_1, l_2, l_3, l_5\}$ is not a time-respective path.



(a) Delivery delay from bus 3029 to 3038 on April 4, 2006

(b) Copies made from bus 3029 to 3038 on April 4, 2006

(c) CDF of delay for all bus pair during the whole trace

Figure 3: Performance of epidemic routing under the trace

versus packet-generation-time plot of a node pair is piecewise linear[1]. As Figure 3(a) shows, the delay versus generatime-time plot is made up of multiple line segments, connected with vertical lines at time instances when a previous path becomes invalid, or when a new path is used from that time on. Similarly, the number-of-copies-made versus generation-time plot and the hop-count versus generation-time plot are also piecewise linear, and more specifically, step functions. Secondly, we observe that the time instances when contacts start (or end) are the time instances when new paths become available (or existing paths become invalid).

Based on the above piecewise linear property, we generate packets at time instances when contacts start and end, and then using the metrics (delay, copies made, hop count) obtained for these packets to obtain these metrics for packets generated at any time. More specifically, for each source-destination pair, a packet was generated respectively at the simulation start time, the starting time $s(l)$, the ending time $e(l)$, and *right after the ending time* $e(l)+$, of each contact.[2]. We then perform trace-driven simulations of epidemic routing for these trace packets.

As an example, Figure 3(a) plots the delivery delay under epidemic routing and direct source-to-destination

[1] Actually, this class of routing schemes include all routing scheme where the forwarding/routing decision at a node does not change in between the two subsequent node-to-node contacts in the network.

[2] The packets generated right after $e(l)$, i.e., $e(l)+$, have a generation time of $e(l)$, but is marked as *trail* packets such that it cannot be sent during contact l.

transmission for packets sent from bus 3029 destined to bus 3038 at any time between 7 A.M. and 7 P.M. on April 4, 2006. We observe a significant difference between the delays achieved by epidemic routing and by direct transmission (i.e., where only the source can deliver a packet directly to a destination). As the two buses have only one contact on this day, the delay under direct transmission is very large. Epidemic routing, however, is able to make use of other buses to relay packets, achieving an average delay of 67.5 minutes. Figure 3(b) plots the number of copies made for a packet generated on the same day for this unicast pair under IMMUNE (where the acknowledgment is not propagated in the network; only destination node can "cure" infected nodes) and VACCINE recovery.

Making use of the piece-wise linear property of the above delay-(copies, and hop count) -versus-generation-time figures, we can evaluate the cumulative distribution function for these performance metrics *assuming packets arrive uniformly randomly to each bus pair, at times uniformly randomly distributed between 7 A.M. to 7 P.M. for all the 55 days in the trace.* For example, Figure 3(c) plots the cumulative distribution function of packet delivery delay for all bus pairs over the whole trace. Our goal is to build a generative model that accurately captures DTN routing performance in terms of the above cumulative distribution functions of delivery delay, copies made and hop counts under epidemic routing.

4 An aggregate model for bus DTN

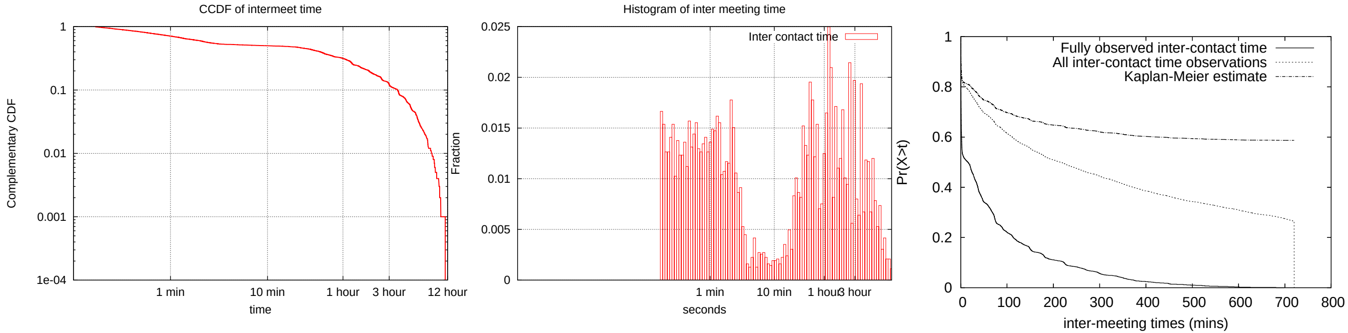
The contact processes between node pairs, and in particular, the inter-contact times between node pairs, determine DTN routing performance. In this section, we characterize and model the bus mobility traces by studying the all-bus-pairs-all-day aggregated inter-contact time. Such approach has been taken by Chaintreau *et al.* [7]. The underlying assumptions made by such approach are *(i)* the contact processes of node pairs are renewal processes, *(ii)* there is no correlation between the contact processes of different node pairs.

In the remainder of this section, we first define the different inter-contact time observations in the trace. We then present the aggregate inter-contact time statistics. Finally, we evaluate a generative model based on the aggregate statistics.

4.1 Censored observations of inter-contact times

Recall that in Section 2.2, we have defined the inter-contact time as the time between two subsequent contacts. For any mobility trace, however, we have different inter-contact time observations. First, there are *fully observed inter-contact time*, measured as the duration of time from the end of a contact to the beginning of the subsequent contact, such as X_2, X_3 in Figure 1. There are also some incomplete observations of inter-contact times. Suppose that we measure the system from 7 A.M. to 7 P.M., for each bus pair, the duration from 7 A.M. to their first observed contact, such as X_1 in Figure 1, is a censored observation. We refer to such an observation as a *start-censored* observation, as we don't know when the inter-contact time starts. Similarly, the duration of time from the last contact between a bus pair to 7 P.M. is also a censored observation, which we refer to as an *end-censored* observation. For the case when two buses have no contacts during this measurement period, we have a *no-meeting* observation with duration given by 12 hr for the bus pair. For such an observation, we do not know the starting or the ending time of the inter-contact time.

To our knowledge, previous studies of mobility traces studied the inter-contact time solely based on fully observed inter-contact times and simply ignored censored observations, with Chen *et al.* [8] as an exception.



(a) CCDF of fully observed inter-contact times

(b) Histogram of fully observed inter-contact times

(c) Empirical CCDFs and KM estimate

Figure 4: Aggregate inter-contact times

Chaintreau *et al.* [7] recognized the effect of finite measurement duration, but did not consider its effect in their characterization of inter-contact time. As longer inter-contact times are more likely to be censored, ignoring censored observations leads to an under-estimation of the inter-contact time distribution, especially when the duration of the measurement is short.

4.2 Aggregate inter-contact time statistics

To study aggregate inter-contact time statistics, we first analyze, for each day, the contact process for each bus pair and obtain censored and fully observed inter-contact times. We then aggregate all the fully observed inter-contact times and censored observations together.

Figures 4(a) and (b) plot the empirical complementary cumulative distribution function (ECCDF) and histogram of the aggregated fully observed inter-contact times, respectively. We observe that the fully observed inter-contact time distribution has two modes, and that there are many short inter-contact times.

The above figures do not suggest an obvious model for the aggregate inter-contact time distribution. Hence we adopt the standard Kaplan-Meier estimator (KM estimator) [20] to estimate the CCDF of the aggregate inter-contact time (also called a survival function), $S(t) := Pr(X > t)$, based on all observations. Suppose there are n distinct fully observed inter-contact times in the sample as follows: $T_1 < T_2 < ... < T_n$, and let $n_i, 1 \leq i \leq n$ be the number of inter-contact times, including both fully observed and censored observations, that are greater than or equal to T_i, and let $d_i, 1 \leq i \leq n$ be the number of inter-contact times of length T_i, then the KM estimator for $S(t)$ is:

$$\hat{S}(t) = \prod_{T_i < t} \frac{n_i - d_i}{n_i}. \tag{1}$$

Eq.(1) is the non-parametric maximum likelihood estimate of $S(t)$.

Figure 4(c) compares the survival function for inter-contact time (i.e., $Pr(X > t)$) estimated by the KM estimator, the ECCDF of fully observed inter-contact times, and the ECCDF of all observations. The results show a very large difference between the CCDF of fully observed inter-contact time and $\hat{S}(t)$. This comparison demonstrates quantitatively the importance of carefully accounting for censored observations when modeling inter-contact times.

4.3 Generative model based on aggregated statistics

In order to accurately model a DTN mobility trace, is it sufficient to model the all-node-pairs aggregate inter-contact time? To answer this question,we compare the routing performance of the original traces to the synthetic tracesgenerated based on the inter-contact time statistics.

To generate a synthetic trace that is comparable to the original trace, we generate traces for the same number of days. For each day, we generate the same number of active buses as in the original trace. The contact process between each bus pair for each day is generated as follows: we draw the time until their first contact (since 7 A.M.) from the observed samples of all the start-censored and no-meeting observations. The subsequent inter-contact times are drawn based on the KM estimate of the conditional distribution of inter-contact time given two buses have contacts on the day, calculated using fully observed inter-contact times and end-censored observations. The contact durations are drawn uniformly and randomly from the aggregate contact duration samples.

We first compare the number of contacts per day in the synthetic trace with that in the original trace. Figure 5 compares the scatter plots of the number of contacts versus the number of active nodes for all the days in the original trace and in the generated trace. It shows that the aggregate model generates a similar total number of contacts per day as the original trace.

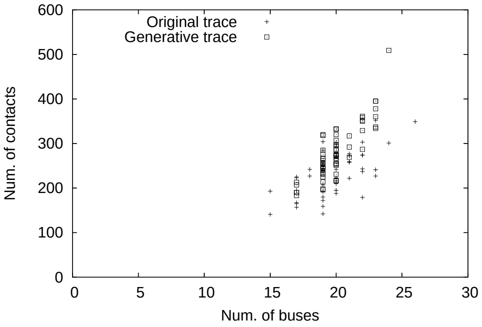
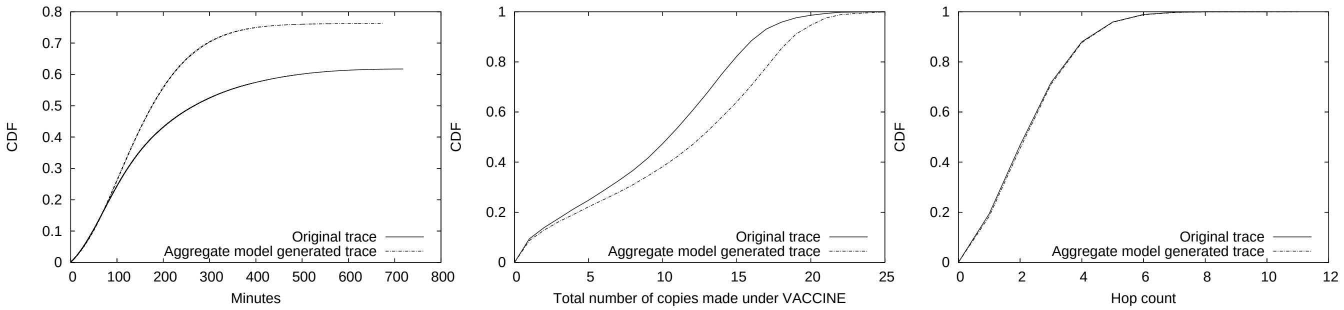


Figure 5: Comparison of no. of active nodes and contacts in aggregated model generated trace and original trace

We then compare epidemic routing performances under the two traces. Figure 6 compares the all-bus-pairs-aggregated CDFs for delivery delay, total copies made in the network, and hop count under the original trace and the generated trace. The results show that many more packets are delivered and fewer copies are made for packets based on the generated trace than on the real trace, although the two traces have a similar number of contacts. (The CDF of the epidemic path hop count, however, is very close.) The reason is that under the aggregate model, contacts are equally distributed to all bus pairs, leading to more balanced connectivities for all buses, which in turn results in more packets being delivered. In fact, we observe that, under original traces, the delivery delays of different bus pairs can differ quite significantly, whereas the generated trace incurs similar performance for different bus pairs. This suggests the need for a finer-grained model to accurately predict DTN routing performance.



(a) Delivery delay CDF (b) No. of copies made CDF (c) Hop count CDF

Figure 6: Comparison of epidemic routing performance under aggregate model generated trace and original trace

5 Modeling Route-level Aggregate Inter-Contact Time

Our study of the aggregate model in the previous section suggests the need for a finer-grain model in order to capture the heterogeneity among different buses. The next question is then: what granularity shall we use to model the mobility trace ?

One approach is to build the finest-grained model possible by characterizing the contacts between individual bus-pairs. This is problematic for two reasons. First, within a day, there are usually just a few contacts between a bus pair; there are simply not enough samples to accurately characterize the pair's contact behavior. Second, each bus is randomly dispatched to a route each day and may change routes during a day, so a bus pair exhibits different meeting behaviors on different days and even during different times of the day. Therefore, one cannot simply aggregate traces from multiple days. For the above reasons, we focus on the contact process between two buses running on certain shift pairs, i.e., shift-pair contact process, rather than the contact process between two physical buses. In the following subsection, we describe the process to construct a shift-pair contact process from the original trace, and we present the route-level aggregate statistics.

5.1 Route-Level Inter-Contact Time Statistics

Recall that for each route in the bus system, there are multiple simultaneous shifts continuously running back and forth on the route. We construct a shift-pair contact process from bus-pair contact processes, making use of the bus dispatching records. Figure 7 illustrates this process. Suppose that we want to generate the contact process between Shift01 and Shift02 (both belong to the SN_SA route). From bus dispatching record, we find that Shift01 (with duration $[t0, t1]$) is served by bus A, while Shift02 is served first by bus B during $[t2, t3]$ and then by bus C during $[t3, t4]$ (as shown by the two middle axes in the diagram). The overlapping time of the two shifts is $[t_s, t_e] = [max(t_0, t_2), min(t_1, t_3)]$. We then insert those contacts between bus A and bus B (shown by top axis) that occur when the buses are running on Shift01 and Shift02 respectively into the (Shift01, Shift02) contact process (shown by the bottom axis), and similarly for bus A and bus C.

In this particular example, our observation of the shift-pair contact process starts at t_s and ends at t_e (i.e., the duration of time that both shifts are actively running). Under our classification of different observations, we have X_1 (X_4) as a start-censored (end-censored) observation for the shift pair, and X_2,X_3 as fully

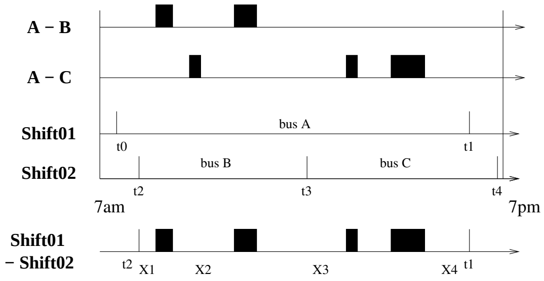


Figure 7: Obtaining (Shift01, Shift02) contact process from original traces using dispatching records. Bus A runs Shift01 during $[t_0, t_1]$, bus B runs Shift02 during $[t_2, t_3]$, bus C runs on Shift02 during $[t_3, t_4]$. For the (Shift01, Shift02) contact process, X_1 (X_4) is a start-censored (end-censored) inter-contact time observation, X_2, X_3 are fully observed inter-contact times.

observed inter-contact times. If we observed no contacts between two shifts, we introduce a no-meeting observation of length $t_e - t_s$.

As we expect different shifts within the same route to exhibit similar contact processes, we aggregate shift-pair inter-contact time observations that belong to the same route pair together to study route-level inter-contact times. For example, Figure 8 plots the histograms of the different observations of the inter-contact time for route pair (SN_SA,SN_SA). Let's first consider the censored observations. We observe the same number of start-censored and end-censored inter-contact times as expected. There are many instances when a pair of buses running on this route have no contacts. The histogram of the fully observed inter-contact times (Figure 8(a)) exhibits interesting periodic behavior and a trend of decreasing probability for longer inter-contact times. There are a large number of small inter-contact times (the first peak in Figure 8(a)). Recall that we have discussed the possible causes for very small inter-contact times in Section 2.2, and we have merged contacts that are less than 60 seconds apart. This figure suggests that there are still many instances of small inter-contact times even after this processing.

The histograms of fully observed inter-contact times for other route-pairs show similar periodic behavior. This suggests that there is interesting structure in the inter-contact times. To better understand the cause of such characteristics, we investigate the deterministic meeting behavior of buses in the next section.

5.2 Understanding Deterministic Meeting Behavior

In this section, we analyze the meeting patterns of two buses running on certain routes based on the assumption that buses operate according to planned schedules and run at constant speed. We define the *inter-meeting time* as the duration of time between when two buses are in transmission range; notice that this is different from *inter-contact time*, which is defined as the duration of time between two subsequent contacts.

We first classify bus routes in our network as either a *linear* or *butterfly-shaped* route. On a linear route, shown in Figure 9(a), bus goes back and forth between two endpoints of the route. On a butterfly-shape route, shown in Figure 9(b), a bus either travels along direction $A \rightarrow B \rightarrow C \rightarrow D \rightarrow E \rightarrow F \rightarrow C \rightarrow D \rightarrow A$ or in the reverse direction.

For two buses running on a same linear route, let the *round trip time* of the route be the time it takes for a bus to travel from an endpoint to another endpoint and then coming back to the starting endpoint, then

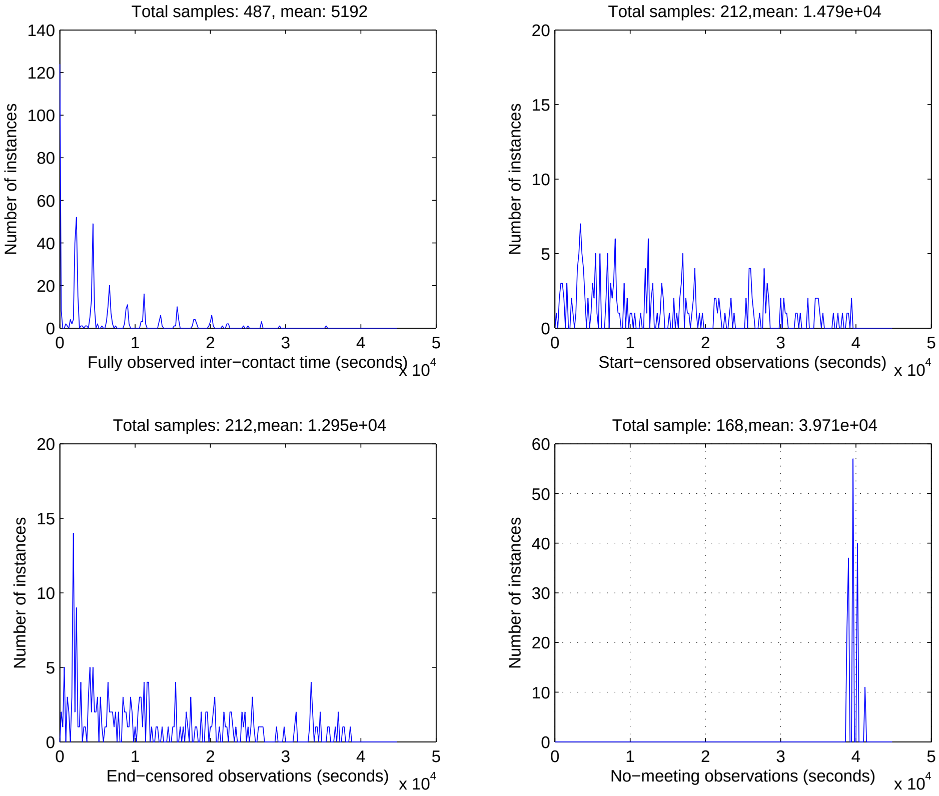
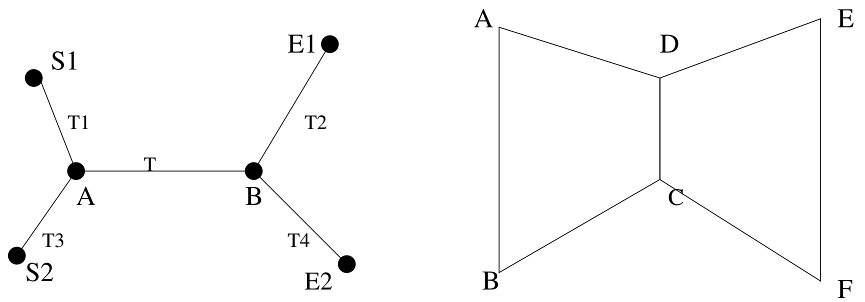


Figure 8: Observations of inter-contact times for SN_SA and SN_SA route pair

they always meet every half round trip time, regardless of the spacing between them.

For a *butterfly-shape route*, let T_l be the travel time for the left loop (ABCDA), and T_r be the travel time for the right loop (CDEFC). The round trip time for the route is given by $T_l + T_r$. It's easy to show that two buses running in opposite directions either follow the inter-meeting time sequence $\{T_l/2, T_l/2, T_r/2, T_r/2, T_l/2, T_l/2, ...\}$ or meet periodically with period $(T_l + T_r)/2$. The latter case occurs when the two buses are spaced so that they do not meet in the joint segment $C - D$. Two buses running in the same direction meet in the $C - D$ segment if their spacing is exactly T_l or T_r. For the butterfly-shape route in our network, i.e., SHUTTLE, we observe that $T_l \approx T_r = T$, therefore a pair of SHUTTLE buses travel in opposite directions have the following inter-meeting time sequences: $\{T, T, ...\}$ or $\{2T, 2T, ...\}$. In addition, SHUTTLE buses running in the same direction very rarely meet, as the buses are scheduled to avoid such meetings.

For two buses running on different routes, we divide the bus routes into smaller segments as needed, and we keep track the time that the buses enter or leave these route segments. If during some time interval, the two buses travel on a same segment in opposite directions, then they will meet each other in the middle of this time interval. For example, for the two linear routes that have overlapping segments (e.g., SN_SA and NA_BR) as shown in Figure 9(a), we consider the the following segments $S_1A, S_2A, AB, BE_1, BE_2$, and let T_1,T_2,T_3,T_4,T be the travel time for each segment. The deterministic inter-meeting times takes up to 5 different values; the inter-meeting time sequence varies depending on the time-phasing of the two buses.



(a) Two linear routes: S1−>A−>B−E1 and S2−>A−>B−E2.

(b) Butterfly Route

Figure 9: Linear route and butterfly-shaped route

5.3 Mean-Restricted Mixture Normal Model

In the previous section, we considered the deterministic meeting sequences between bus pairs, ignoring random influence such as varying traffic and bus-operation conditions. We found that two buses running on a specific route pair had a fixed meeting sequence that is made up a number of inter-meeting times $T_{bi}, i = 1, 2, 3, ..., k$.

In reality, due to varying traffic conditions, bus speeds and other considerations, the inter-meeting time of buses is not constant, but rather a random variable that we can model as a normal distribution with mean T_{bi} and a certain variance. Furthermore, when two buses are in transmission range of each other, they are not always able to associate and transfer data, due to high bus speed, or because one of the buses is already in contact with a fixed access point. As a result, a data transfer can occur at the l-th physical meeting since the last contact ($l = 1, 2, 3, ...$). This means that an inter-contact time is made up of l inter-meeting times. As each inter-meeting time can be modeled as a normal random variable with mean given by $T_{bi}, i = 1, 2, 3, ..., k$, the inter-contact times can be modeled as the sum of l such normal random variables where $l = 1, 2, 3, ...$

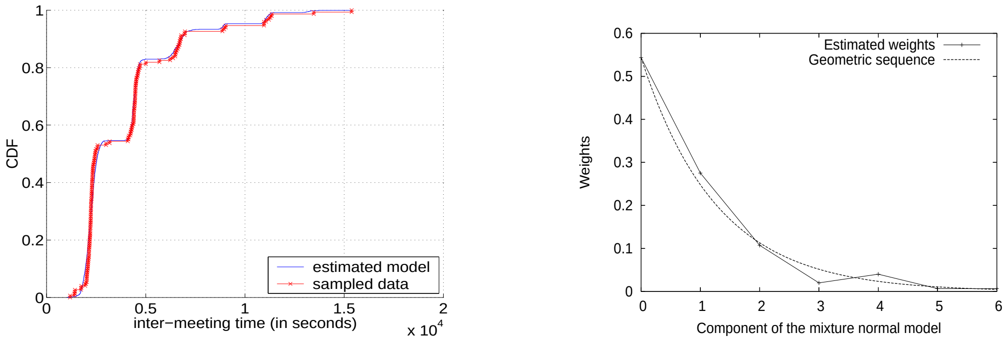
For the case where there is a single inter-meeting time between a bus pair running on the route-pair, e.g., (SN_SA, SN_SA) and (NA_BR, NA_BR), or when the inter-meeting times are multiples of a single base value, e.g., (SHUTTLE,SHUTTLE) route pair, we propose the following mixture normal model for the inter-contact times:

$$f_{SM}(x) = \sum_{i=1}^{G} w_i f_N(x|i\mu, \sigma^2), \tag{2}$$

where f_N represents the PDF of the normal distribution parameterized by mean $i\mu$ and variance σ^2, μ corresponds to the base inter-meeting time T_b, σ^2 is the common variance for all normal components, the weights w_i depend on the specific inter-meeting time sequence for the route-pair, and we have $\sum_{i=1}^{G} w_i = 1$.

We derive an Expectation Maximization algorithm [5] to estimate the model parameters from fully observed inter-contact times[3]. The detail of the EM algorithm is given in Appendix A. As this model, and the model in the next section, focus on the periodicity of inter-contact times, we have excluded the short inter-contact time observations when applying the model. We applied the model to study the (SN_SA,SN_SA)

[3] Censored observations are not considered here, as we propose a model that is more appropriate for taking into account censored data in Section 5.4.



(a) CDFs of data and estimated model

(b) Estimated weights and geometric seq.

Figure 10: Model fitting result for mean-restricted mixture normal models for SN_SA and SN_SA data

data set, and compared the empirical CDF of fully observed inter-contact time (with short inter-contact time removed) with that of the estimated model in Figure 10(a). We find that they match very well.

The above model has incorporated the periodicity by setting the means of the normals to be multiples of a single base value. From the original data (e.g., Figure 8), we also observe a geometric trend in the heights of the different normal components. In Figure 10(b), we plot the estimated weights, and we find that they match quite closely with the curve of the geometric sequence $p^{i-1}(1-p)$, with $p = 1 - w_0$. Actually, if we assume there is a fixed probability that two buses fail to set up a contact when they meet, then we have $w_i = p^{i-1}(1-p)$, where i is the number of meetings until a successful contact.

As for the case when there are multiple inter-meeting times between a bus pair running on a route pair, such as SN_SA and NA_BR route pair, one could consider a mixture of normals with the means set to different linear combinations of the basic inter-meeting times. As we don't have enough data samples for such route pairs in our network, i.e., (SN_SA, NA_BR), (SN_SA, SHUTTLE) and (NA_BR, SHUTTLE), we leave the modeling of them for future work.

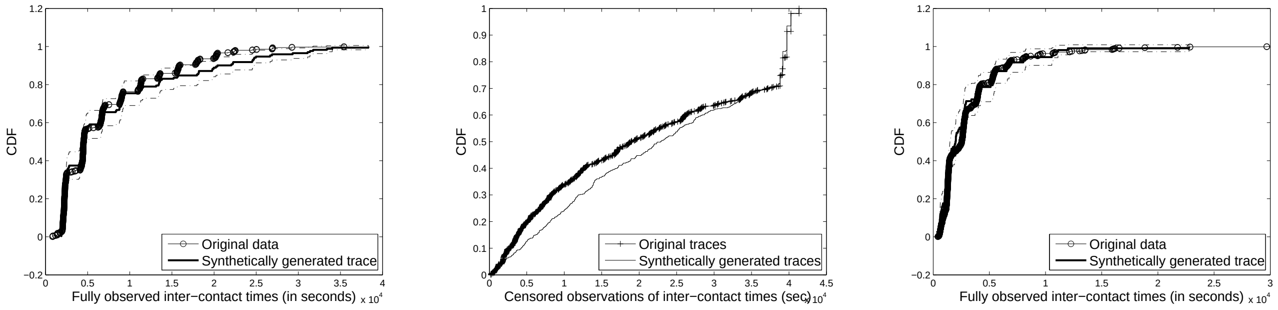
5.4 Mean-Weight-Restricted Mixture Normal Model

The model proposed in the previous section has incorporated our knowledge about the deterministic meeting sequences of the route pair, but still involves parameters that have no clear physical interpretations, i.e., the weights and the number of components. Furthermore, it's not clear how to take into account censored observations when estimating model parameters. Nevertheless, our analysis of the weights as estimated by the model parameter estimation algorithm has suggested the following models that explicitly model the probability of failing to set up contact when buses are in range.

One-Base-Mean Model. For bus pairs with a single inter-meeting time, one can use the following model to characterize their inter-contact times:

$$f_{GEO\_1P\_1BM}(x) = \sum_{i=1}^{\infty} p^{i-1}(1-p) f_N(x|i\mu, \sigma^2), \tag{3}$$

where p is the probability that two nodes in transmission range fail to establish a contact, μ corresponds to the base inter-meeting time, and a single variance σ^2 is used for all normals as buses tend to keep to their schedules, so the variance does not add up. As there is a single base inter-meeting time μ, we refer it as



(a) Fully observed inter-contact times for (SN_SA,SN_SA)

(b) Censored observations for (SN_SA, SN_SA)

(c) Fully observed inter-contact time (SHUTTLE,SHUTTLE)

Figure 11: Model fitting results for mean-weight-restricted mixture normal model

one-base-mean model.

In Section 4.2, we have demonstrated that one needs to consider both fully observed inter-contact times and censored observations in order to correctly characterize the inter-contact time. Recall that we have used Kaplan-Meier estimator when we do not assume a model for inter-contact times. We now discuss how to account for censored observations when estimating parameters for the above model.

To consider censored data in the model parameter estimation, we first need to understand how the censored observations relate to the inter-contact times. Let's denote the PDF (Probability Distribution Function) and CDF of the inter-contact time as $f_X(x)$ and $F_X(x)$ respectively, and denote the mean of the inter-contact time as $E[X]$. We assume that the time we start to observe a shift-pair, i.e., the time that two buses enter the routes (t_s in Figure 7), is a random incidence into an inter-contact time interval. As a result, the time until we see the next contact, i.e., the start-censored observation is the residual lifetime [23] following a PDF given by $f_Y(x) = \frac{1-F_X(x)}{EX}$. If we observe no meeting, i.e., a no-meeting censored observation of length $t_e - t_s$, this means that we observe a residual life time that is longer than $t_e - t_s$ (the probability of which is given by $1 - F_Y(t_e - t_s)$, where $F_Y(x)$ is the CDF for $f_Y(x)$.). We further assume that all inter-contact times are equally likely to be cut off in the end. Thus an end-censored observation of value y means that a random inter-contact time has value larger than y, the probability of which is then $1 - F_X(y)$. Based on the above analysis of the censored data, we derive an EM algorithm to estimate the parameters p, μ, σ in model (3) from empirical data (details are given in Appendix B).

It turns out that the above model doesn't provide a good fit to the (SN_SA,SN_SA) data set. A careful examination of the traces and the bus schedule reveals that some shift pairs have fewer contacts than other shift pairs. This is mainly due to the fact that different shift pairs meet at different points within the route segment, some meet at high speeds, others meet at more congested downtown areas at low speeds. When buses traveling at high speed come into transmission range, there is shorter duration of time for them to set up connection and transfer data, which means a higher failure probability in setting up a contact. Based on these observations, we extend the above model to account for such factors:

$$f_{GEO\_MP\_1BM}(x) = \sum_{i=1}^{C} w_i \sum_{l=1}^{\infty} p_i^{l-1}(1-p_i) f_N(x|l\mu, \sigma^2), \tag{4}$$

where C is the number of components, and w_i specifies the fraction of bus pairs that have failure probability given by p_i.

Similar to model (3), we derive an EM algorithm to estimate the parameters for the above model, using all observations. We then use model (4) with $C = 2$ to model (SN_SA, SN_SA) data set, and then generate synthetic traces based on the estimated model. Figure 11(a) and (b) respectively compare the CCDF of the model-generated fully observed inter-contact time and censored observations with those in the original traces. We observe that for the fully observed inter-contact times, the original data fall within the 95% confidence interval of the model. As to the censored observation, the match is less good. We believe that this is due to the fact that there are other failure conditions that haven't been taken into account in the model, such as bus hardware failure or hardware being turned off for certain duration.

Two-Base-Mean Model. Recall from our deterministic analysis of the SHUTTLE route that *(i)* a bus pair running on SHUTTLE in the opposite direction either meet every half round trip time T or every round trip time $2T$; *(ii)* a bus pair running in the same direction very rarely meet with each other. Based on this knowledge, we propose the following model for the inter-contact time for a pair of SHUTTLE buses running in the opposite directions:

$$f_{GEO\_2BM}(x) = \sum_{i=1}^{2} w_i \sum_{l=1}^{\infty} p_i^{l-1}(1-p_i) f_N(x|li\mu, \sigma^2). \tag{5}$$

where p is the probability that two buses in transmission range fail to establish a contact, $\mu, 2\mu$ correspond to the base inter-meeting times T and $2T$, and a single variance σ^2. We refer this model as two-base-mean model as there are two base inter-meeting times, μ and 2μ.

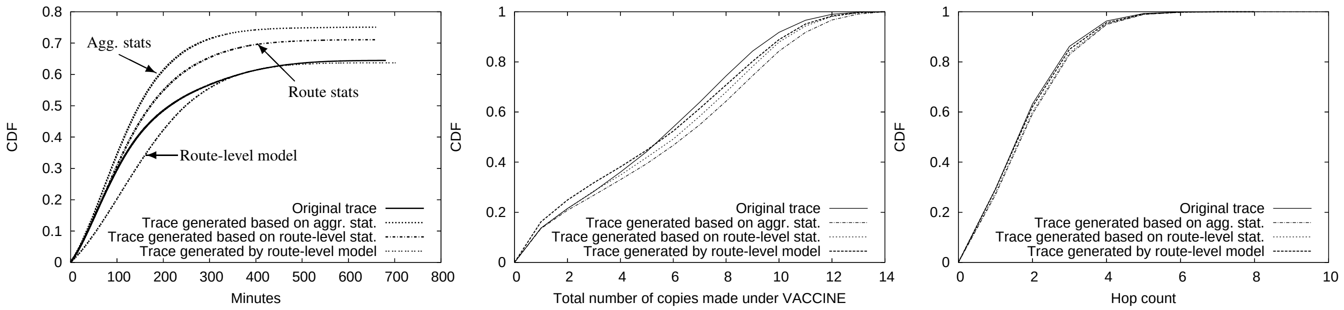
Similar to the One-Base-Mean model, we develop an EM algorithm to estimate the parameters for the above model from the fully observed data and censored data. We apply the model to (SHUTTLE, SHUTTLE) dataset, and the results are plotted in Figure 11(c), which shows a good fit of our model to the empirical data.

5.5 Model Comparison

In this section, we compare three models, i.e., the model based on all-shift-pair aggregate statistics, the model based on route-level statistics, and the route-level model described in the last section, with a focus on their accuracy in capturing epidemic routing performance of the original trace.

We first process the original traces to include only buses running on the three routes that we have been focusing on, and we analyze this thinned trace to obtain aggregated inter-contact time statistics and route-level aggregated statistics. We then generate three synthetic traces. The first synthetic trace is generated based on all-shift-pair aggregated statistics, using the procedure described in Section 4.3; the second synthetic trace is generated based on the route-level statistics in a similar way; the third synthetic trace is generated using the route-level inter-contact models that we developed based on the route-level statistics in the last section, combined with route-level statistics for route-pairs that we don't have a model for. Last, we simulate epidemic routing respectively over the thinned original trace and the three synthetic traces. Figure 12 compares routing performance in terms of delivery delay, copies made, and the hop count of minimal delay paths under the four traces. We observe that under the four traces, the difference in delivery delay is the largest, followed by copies made. All traces have similar hop count CDF.

Similar to what we have observed in Figure 6 in Section 4.3, the trace generated by aggregate model exhibits significantly different performance compared to the original trace (i.e., the trace of inter-contact times empirically observed in the operational bus network). The trace based on route-level statistics, which



(a) CDF for delivery delay (b) CDF for no. of copies made (c) CDF of hop count

Figure 12: Comparison of epidemic routing performance under original trace and synthetic traces generated by three different models, respectively

is able to capture the heterogeneity among different bus routes, exhibits epidemic routing performance closer to the original trace.

Now let's focus on the the route-level model generated trace. We find that under this trace, all performance metrics are closer to the original trace than those of the previous two traces. In particular, under this trace, the average delivery delay is 15.8% larger than that under the original trace; and the packet delivery ratio is 0.75% less than that under the original trace. We think that the larger delivery delay and slightly smaller delivery ratio are due to the fact that our route-level model does not capture those short inter-contact times, and therefore it generates fewer contacts than the original ones. Nevertheless, as the model captures longer inter-contact times accurately, it's able to predict the longer time range delivery performance well.

As the route-level models are developed based on route-level statistics, it's somehow surprising that the prediction performance of the former is better than the latter. Our explanation is that in the route-level model we developed, we have treated the SHUTTLE bus pairs traveling in the same direction, and those traveling in the opposite direction differently, whereas in route-level statistics, we treated them together in SHUTTLE-SHUTTLE route pair. A related comment is that we expect that the granularity between shift- and route-level might be able to capture different meeting behavior exists between different shift-pairs within a route, and therefore will achieve a good balance between model complexity and prediction performance.

6 Related Work

Many previous works have proposed mobility models [14, 31, 17, 22, 25, 15] for Wireless LANs or Vehicular Ad-Hoc networks, with some of them based on real mobility traces [31, 22]. Our work differs significantly from the above because we model the contact process between node pairs and in particular, the inter-contact time distribution between node pairs, as contact opportunity frequency and duration are the main determinants of DTN routing performance. Moreover, our traces are generally longer and more fine-grained than those used in previous studies.

Due to the lack of large scale DTN deployment, traces collected from campus WLANs such as UCSD [2] and Dartmouth College [13] of access point (AP) association records for laptops and PDAs were sometimes adapted to support DTN research. This is based on the assumption that mobile devices associating with the same AP would also be in transmission range of each other. Recently, several projects have collected traces from real DTNs, including the University of Toronto trace [29], iMote traces by Haggle project [26, 27]. The Reality Mining project at MIT [1, 11] has also made available node-to-node contact traces of mobile phones carried by students and faculty.

A couple of works have analyzed mobility traces and studied the impact on routing schemes. For example, Hsu and Helmy [16] also used the Dartmouth traces [13] to study encounter-based broadcasting. They studied the trace induced *encounter relationship graph* (where a pair of nodes is connected with an edge if they ever meet each other), found it exhibits a small world property, and showed that encounter-based forwarding is robust to selfish node behaviors. For both works, the traces used are WLAN traces, rather than a real trace collected from a DTN, and the approximation with the same access point is used to infer contacts between a pair of devices. Another example is the works by Su *et al.* [30, 29] that studied DTN traces collected from a network of 20 students carrying PDAs with bluetooth radio. They studied direct contact and multi-hops paths between node pairs, and used trace driven simulation of epidemic routing and link-state routing to characterize the trade-off between delay and replication.

We next review several recent works that characterize DTN mobility traces through studying the inter-contact times between node pairs, similar to our works.

First, Chaintreau *et al.* [7] characterized the all-node-pairs aggregate inter-contact times from the UCSD trace [2], Dartmouth trace [13], Toronto trace [29] and three iMote traces, all of pedestrians carrying wireless devices. They observed an approximately power-law distribution of inter-contact times, with a power law index less than 1. Based on this observation, the authors proposed a simplified stationary *i.i.d.* contact model with power-law distributed inter-contact times, and they analytically studied the performance of different forwarding algorithms under such a model. The question of whether an aggregate model is sufficient for predicting DTN routing performance was not addressed. Their study did not account for censored observations in the characterization of aggregate inter-contact times. Furthermore, the claim that inter-meeting time follows power-law distribution needs more careful examination.

In their analysis of the UCSD and Dartmouth trace [2, 13], Chen *et al.* [8] took into account censored observations in their characterization of the aggregate inter-contact times, and proposed a censorship removal algorithm.

Karaginnis *et al.* [21] analyzed human mobility traces including the UCSD trace [2], the MIT Reality Mining trace [1, 11], the iMote traces [26, 27], and personal vehicular GPS dataset [24]. They found that for the aggregate inter-contact times of all traces, there is a characteristic time, in the order of half of day, beyond which the distribution decays exponentially; up to this value, the distribution follows a power law. This is in contrast to previous hypothesis of power-law distribution by Chaintreau *et al.*, and suggests the prediction therein on routing schemes performance may be overly pessimistic. They also demonstrated that simple synthetic models can feature the above dichotomy, showing that simple synthetic models might be good for capturing human beings's mobility trace. Through further analysis of the trace, the author demonstrated that the dichotomy can be explained by the underlying returning time to favorite home of the mobile nodes. In addition to the above analysis, the author further explored the spatial and temporal heterogeneity within the trace. We comment that they have not considered censored observations, and therefore the conclusions about exponential decay can be an under-estimate of the actual inter-contact times.

A recent work by Conan *et al.* [9] studied the Dartmouth traces, iMote traces and the Reality Mining traces. Other than focusing on aggregate inter-contact time distribution, they studied pair-wise inter-contact times and found that log-normal distribution fits the largest fraction of data. They also provided a potential explanation to the approximate power-law distribution observed in the aggregate inter-contact times: when pair-wise inter-contact times follow exponential distribution with different rate, one can gain power-law phenomenon in the aggregate statistics. Finally, for a DTN with exponential pair-wise inter-contact times (with different rate for each pair), an opportunistic single-copy routing scheme (that minimize expected delay) is proposed and evaluated against other schemes.

7 Summary

In this work, we have studied mobility traces taken from UMass DieselNet, with the goal of building a generative model that can capture aspects of mobility (specifically inter-contact times) at the right level of granularity. The model is generative in that it can be used to generate synthetic traces to drive a trace-driven simulation. Although this model is derived from mobility traces collected from a specific bus-based network, we expect such a model is applicable to other transport based networks that follow certain periodic schedules. Further work is needed to validate the model using mobility traces collected from different networks once they become available.

As the first careful study of a fielded system, the model is of interest in its own right, as they revealed structure that was hidden at the aggregate level — structure that can influence DTN performance. Indeed, using a trace-driven simulation of epidemic routing, we showed that this finer grained route-level model of inter-contact times predicts performance much more accurately than the coarser-grained aggregated all-bus-pairs model. This suggests that one must take care in choosing the right level of model granularity when modeling mobility-related measures such as inter-contact times, in DTN networks. Determining the appropriate granularity of models is both a difficult and a deep problem. At one extreme we can use a movement model, such as Brownian motion with parameters that are chosen to correspond to the parameters of the aggregate inter-contact time distribution obtained from a trace. At another extreme we can devise a model that accounts for the physics of the underlying system such as described above for our bus-based DTN network.

Our ongoing work includes the understanding and modeling of short inter-contact times. Our future research will focus on identifying the level of abstraction needed to produce good models, where goodness refers to how well generated traces statistically match collected mobility traces and how well models predicting the behavior of different information dissemination algorithms. We will also focus on developing techniques for teasing out the physical structure from a trace (such as the underlying periodic behavior in the inter-contact times) in the absence of domain knowledge.

8 Acknowledgement

This research was supported in part by the National Science Foundation under award number ANI-0325868 and CNS-0519881. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies. This research is also continuing through participation in the International Technology Alliance sponsored by the U.S. Army Research Laboratory and the U.K. Ministry of Defense under Agreement Number W911NF-06-3-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defense or the U.K. Government. The U.S. and U.K. Govenments

are authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright notation hereon.

A EM algorithm for mean-restricted mixture normal model

In this section, we outline the EM algorithm for the mean-restricted mixture normal model Eq.(2) proposed in Section 5.3.

Suppose that we have N fully observed inter-contact times: $x_i, i = 1, 2, ..., N$. The following EM algorithm is used to iteratively estimate μ, σ^2, and $w_l, \forall l \in 1, 2, ..., G$ from these fully observed inter-contact times:

$$w_{t+1,l} = (1/N) \sum_{i=1}^{N} p(l|x_i, \Theta_t) \tag{6}$$

$$\mu_{t+1} = \frac{\sum_{l=1}^{G} l \sum_{i=1}^{N} x_i p(l|x_i, \Theta_t)}{\sum_{l=1}^{G} l^2 \sum_{i=1}^{N} p(l|x_i, \Theta_t)} \tag{7}$$

$$(\sigma^2)_{t+1} = \frac{\sum_{l=1}^{G} \sum_{i=1}^{N} (x_i - l\mu_{t+1})^2 p(l|x_i, \Theta_t)}{\sum_{l=1}^{G} \sum_{i=1}^{n} p(l|x_i, \Theta_t)} \tag{8}$$

where $t = 1, 2, ...$ is the iterative step, $\Theta_t = \{\mu_t, (\sigma^2)_t, w_{l,t}, l = 1, 2, ..., G\}$ is the current estimate of the parameters, and $p(l|x_i, \Theta_t)$ is the probability that the random sample x_i comes from component l given the model parameter vector Θ_t. By Bayes' Rule, we have:

$$p(l|x_i, \Theta_t) = \frac{p(l, x_i|\Theta_t)}{p(x_i|\Theta_t)} = \frac{p_t^{l-1}(1 - p_t) f_N(x_i, lu_t, \sigma_t^2)}{\Sigma_{j=1}^{\infty} p_t^{j-1}(1 - p_t) f_N(x_i, j\mu_t, \sigma_t^2)},$$

here the function $f_N(x, \mu, \sigma^2)$ represents the PDF of normal distribution with mean μ and variance σ^2.

B EM algorithm for One-Base-Mean Model

In this section, we first outline the derivation of Expectation-Maximization algorithm for the One-Base-Mean model (Eq.(3) in Section 5.4), and then give the EM algorithm for model given by Eq.(4) and Two-Base-Mean model.

One-Base-Mean Model. We start by outlineing the derivation of EM algorithm for the following model:

$$f_{GEO\_1P\_1BM}(x) = \sum_{i=1}^{\infty} p^{i-1}(1 - p) f_N(x|i\mu, \sigma^2).$$

Our goal is to derive the maximum likelihood estimate for the model parameters $\Theta = (p, \mu, sigma^2)$, based all observations of the inter-contact times. Assume that we have N fully observed inter-contact time, x_i, $(i = 1, ..., N)$, N_s start-censored inter-contact time observations, S_i,$(i = 1, ..., N_s)$, N_e end-censored inter-contact time observations, E_i,$(i = 1, ..., N_e)$, and N_n no-meeting observations, N_i,$(i = 1, ..., N_n)$.

We assume that the fully inter-contact time x_i's are independent and identically distributed (i.i.d.) with PDF given by above model, and the distributions of the censored observations relate to the above model as to our discussion in Section 5.4. We denote the whole data set as:

$$X = (x_1, ..., x_N, S_1, ..., S_{N_s}, E_1, ..., E_{N_e}, N_1, ..., N_{N_n}).$$

Due to the complex form of the model, we resort to EM algorithm to obtain the maximum likelihood estimate.

First, we introduce hidden variables for each observations (i.e., samples) of the inter-contact time, representing the number of physical meetings within that inter-contact time. We denote the whole set of hidden variables as Y.

Next, in the Expectation-step, we derive $p(l|x_i, \Theta_t)$ ($p(l|S_i, \Theta_t)$, $p(l|E_i, \Theta_t)$, $p(l|N_i, \Theta_t)$), i.e., the distribution of the hidden variable (the number of physical meetings within the inter-contact time), given the fully observed data (started-censored data, end-censored data, no-meeting observations) and the current estimates of model parameters, $\Theta_t = (p_t, \mu_t, \sigma_t^2)$.

(1) **Fully observed inter-contact times**. By Bayes' Rule, we have:

$$\begin{aligned} p(l|x_i, \Theta_t) \qquad &= \frac{p(l, x_i|\Theta_t)}{p(x_i|\Theta_t)} \\ &= \frac{p_t^{l-1}(1-p_t) f_N(x_i, lu_t, \sigma_t^2)}{\Sigma_{j=1}^{\infty} p_t^{j-1}(1-p_t) f_N(x_i, j\mu_t, \sigma_t^2)} \end{aligned}$$

for $i = 1, ..., N$, $l = 1, 2, ...$.

(2) **Start-censored observations**. As discussed in Section 5.4, we assume such observation is the residual lifetime, and its PDF is given by $f_Y(x) = \frac{1-F_X(x)}{EX}$, where $F_X(x), EX$ is the CDF and mean of the inter-contact time respectively. We have:

$$g_l(S_i) := p(S_i|l, \Theta_t) = \int_{S_i}^{\infty} f_N(x, l\mu_t, \sigma_t^2) dx / l\mu_t. \tag{9}$$

The conditional distribution for the number of physical meetings given the start-censored observation is given by:

$$\begin{aligned} p(l|S_i, \Theta_t) \qquad &= \frac{p(l, S_i|\Theta_t)}{p(S_i|\Theta_t)} \\ &= \frac{\int_{S_i}^{\infty} p_t^{l-1}(1-p_t) f_N(x, l\mu_t, \sigma_t^2) dx / l\mu_t}{\sum_{j=1}^{\infty} p_t^{j-1}(1-p_t) \int_{S_i}^{\infty} f_N(x, j\mu_t, \sigma_t^2) dx / j\mu_t} \end{aligned}$$

for $S_i, i = 1, ..., N_s, l = 1, 2, ...$.

(3) **No-meeting observations**. Assume the CDF corresponding to the PDF $g_l(x)$, i.e., Eq.(9), is $G_l(x)$, i.e., $G_l(x) = \int_{-\infty}^{x} g_l(y) dy$. We have

$$p(l|N_i, \Theta_t) = \frac{p(l, N_i|\Theta_t)}{p(N_i|\Theta_t)} = \frac{p_t^{l-1}(1-p_t)(1-G_l(N_i))}{\sum_{j=1}^{\infty} p_t^{j-1}(1-p_t)(1-G_j(N_i))},$$

for $i = 1, ..., N_n, l = 1, 2, ...$

(4) **End-censored observations**. We have

$$\begin{aligned} p(l|E_i, \Theta_t) &= \frac{p(l, E_i|\Theta_t)}{p(E_i|\Theta_t)} \\ &= \frac{\int_{E_i}^{\infty} p_t^{l-1}(1-p_t) f_N(x, l\mu_t, \sigma_t^2)dx}{\sum_{j=1}^{\infty} p_t^{j-1}(1-p_t) \int_{E_i}^{\infty} f_N(x, j\mu_t, \sigma_t^2)dx} \end{aligned}$$

for $i = 1, ..., Ne, l = 1, 2, ...$

Last, in the Maximization-step, we derive the expectation of log complete data likelihood function, conditioned on the observations and the current estimates of model parameters, as follows:

$$Q(\Theta, \Theta_t) := E[log(P(X, Y|\Theta)|X, \Theta_t],$$

and obtain new estimates for the model parameters by setting:

$$\Theta_{t+1} = \text{argmax}_{\Theta} Q(\Theta, \Theta_t).$$

Omitting the derivations (which is similar to that in [5]), we get the following updating rule:

$$\begin{aligned} p_{t+1} &= num_p/denom_p \\ \mu_{t+1} &= \frac{\sum_{i=1}^{N}\sum_{l=1}^{\infty} x_i l p(l|x_i, \Theta_t)}{\sum_{i=1}^{N}\sum_{l=1}^{\infty} l^2 p(l|x_i, \Theta_t)} \\ (\sigma^2)_{t+1} &= \frac{\sum_{i=1}^{N}\sum_{l=1}^{\infty}(x_i - l\mu_{t+1})^2 p(l|x_i, \Theta^t)}{\sum_{i=1}^{N}\sum_{l=1}^{\infty} p(l|x_i, \Theta_t)} \end{aligned}$$

where $t = 1, 2, ..$ is the iterative step, and

$$\begin{aligned} num_p &= \sum_{i=1}^{N}\sum_{l=1}^{\infty}(l-1)p(l|x_i, \Theta_t) + \sum_{i=1}^{N_s}\sum_{l=1}^{\infty}(l-1)p(l|S_i, \Theta_t) \\ &\quad + \sum_{i=1}^{N_e}\sum_{l=1}^{\infty}(l-1)p(l|E_i, \Theta_t) + \sum_{i=1}^{N_n}\sum_{l=1}^{\infty}(l-1)p(l|N_i, \Theta_t) \\ denom_p &= \sum_{i=1}^{N}\sum_{l=1}^{\infty} lp(l|x_i, \Theta_t) + \sum_{i=1}^{Ns}\sum_{l=1}^{\infty} lp(l|S_i, \Theta_t) \\ &\quad + \sum_{i=1}^{Ne}\sum_{l=1}^{\infty} lp(l|E_i, \Theta_t) + \sum_{i=1}^{Nn}\sum_{l=1}^{\infty} lp(l|N_i, \Theta_t) \end{aligned}$$

Note that we assume the censored observations only affect the estimate of p, and ignored them while updating μ, σ^2.

One-Base-Mean-Multi-FailureProb Model. Here we outline the EM algorithm for the model Eq.(4) as follows:

$$f_{GEO\_MP\_1BM}(x) = \sum_{i=1}^{C} w_i \sum_{l=1}^{\infty} p_i^{l-1}(1-p_i) f_N(x|l\mu, \sigma^2).$$

To derive EM algorithm for this model, we consider the following two hidden variables for each data samples. One hidden variable, denoted as $c, 1 \leq c \leq C$, designates the failure probability (p_c) governed the observed inter-contact time; another hidden variable, denoted as $l, 1 \leq l \leq \infty$, is the number of physical inter-meeting times within the observed sample.

We have the following updating rules, here $\Theta = (w_1, ..., w_C, p_1, ..., p_C, \mu, \sigma^2)$.

$$\begin{aligned}
(w_j)_{t+1} &= \frac{1}{N}\sum_{i=1}^{N} p(c=j|x_i,\Theta_t) + \frac{1}{N_s}\sum_{i=1}^{N_s} p(c=j|S_i,\Theta_t) \\
&\quad + \frac{1}{N_e}\sum_{i=1}^{N_s} p(c=j|E_i,\Theta_t) + \frac{1}{N_n}\sum_{i=1}^{N_s} p(c=j|N_i,\Theta_t) \\
\mu_{t+1} &= \frac{\sum_{i=1}^{N}\sum_{l=1}^{\infty} x_i l p(l|x_i,\Theta_t)}{\sum_{i=1}^{N}\sum_{l=1}^{\infty} l^2 p(l|x_i,\Theta_t)} \\
(\sigma^2)_{t+1} &= \frac{\sum_{i=1}^{N}\sum_{l=1}^{\infty}(x_i - l\mu_{t+1})^2 p(l|x_i,\Theta^t)}{\sum_{i=1}^{N}\sum_{l=1}^{\infty} p(l|x_i,\Theta_t)} \\
(p_j)_{t+1} &= num_{p_j}/denom_{p_j}
\end{aligned}$$

where $j = 1, ..., C$. The updating rules for $p_j, j = 1, ..., C$, μ and σ^2 are the same as those of the previous model, with the only difference is the evaluation of the conditional distribution of hidden variables given the data and the current estimates of model parameters. By Bay's rule, we have:

$$\begin{aligned}
p(c,l|x_i,\Theta_t) &= \frac{p(x_i,c,l|\Theta_t)}{p(x_i|\Theta_t)} \\
&= \frac{(w_c)_t (p_c)_t^{l-1}(1-(p_c)_t) f_N(x_i, l\mu_t, \sigma_t^2)}{\sum_{j=1}^{2}\sum_{k=1}^{\infty}(w_j)_t (p_j)_t^{k-1}(1-(p_j)_t) f_N(x_i|jk\mu_t, \sigma_t^2)},
\end{aligned}$$

and

$$p(c|x_i,\Theta_t) = \sum_{l=1}^{\infty} p(c,l|x_i,\Theta_t),$$

$$p(l|x_i,\Theta_t) = \sum_{c=1}^{C} p(c,l|x_i,\Theta_t).$$

We have:

$$num_{p_j} = \sum_{i=1}^{N}\sum_{l=1}^{\infty}(l-1)p(c=j,l|x_i,\Theta_t) + \sum_{i=1}^{N_s}\sum_{l=1}^{\infty}(l-1)p(c=j,l|S_i,\Theta_t)$$

$$
\begin{aligned}
& +\sum_{i=1}^{N_e}\sum_{l=1}^{\infty}(l-1)p(c=j,l|E_i,\Theta_t)+\sum_{i=1}^{N_n}\sum_{l=1}^{\infty}(l-1)p(c=j,l|N_i,\Theta_t) \\
denom_{p_j} &= \sum_{i=1}^{N}\sum_{l=1}^{\infty} lp(c=j,l|x_i,\Theta_t)+\sum_{i=1}^{Ns}\sum_{l=1}^{\infty} lp(c=j,l|S_i,\Theta_t) \\
& +\sum_{i=1}^{Ne}\sum_{l=1}^{\infty} lp(c=j,l|E_i,\Theta_t)+\sum_{i=1}^{Nn}\sum_{l=1}^{\infty} lp(c=j,l|N_i,\Theta_t)
\end{aligned}
$$

Similar to the One-Base-Mean model, we derive the conditional distribution of the hidden variables given the censored observations.

Two-Base-Mean Model. We now briefly describe the EM algorithm used for the following model:

$$
f_{GEO\_2BM}(x)=\sum_{i=1}^{2} w_i \sum_{l=1}^{\infty} p_i^{l-1}(1-p_i) f_N(x|li\mu,\sigma^2).
$$

Here we have $\Theta=(w_1,w_2,p_1,p_2,\mu,\sigma^2)$. The derivation of the EM algorithm is similar to the One-base-mean model, except that for this model, we introduce two random variables for each data sample (i.e., observation): one, c denotes the class of the inter-contact time, another, l, denotes the number of physical inter-meeting times within the observed inter-contact time. We evaluate the conditional distribution of the hidden variables as follows:

$$
\begin{aligned}
p(c,l|x_i,\Theta_t) &= \frac{p(x_i,c,l|\Theta_t)}{p(x_i|\Theta_t)} \\
&= \frac{(w_c)_t(p_c)_t^{l-1}(1-(p_c)_t)f_N(x_i,l\mu_t,\sigma_t^2)}{\sum_{j=1}^{2}\sum_{k=1}^{\infty}(w_j)_t(p_j)_t^{k-1}(1-(p_j)_t)f_N(x_i|jk\mu_t,\sigma_t^2)}
\end{aligned}
$$

Similarly, we evaluate the conditional distrbution of the hidden variables for the censored observations.

The updating rules are as follows:

$$
\begin{aligned}
(w_j)_t &= \sum_{l=1}^{\infty}(\sum_{i=1}^{N} p(j,l|x_i,\Theta_t)+\sum_{i=1}^{N_s} p(j,l|S_i,\Theta_t)+\sum_{i=i}^{N_e} p(j,l|E_i,\Theta_t)+\sum_{i=1}^{N_n} p(j,l|N_i,\Theta_t), \\
&\quad \text{for } j=1,...,C \\
(p_j)_t &= \frac{\sum_{l=1}^{\infty}\sum_{i=1}^{N}(l-1)p(j,l|x_i,\Theta_t)}{\sum_{1=1}^{\infty}\sum_{i=1}^{N} lp(j,l|x_i,\Theta_t)}, \text{ for } j=1,...,C \\
\mu_t &= \frac{\sum_{j=1}^{C}\sum_{l=1}^{\infty}\sum_{i=1}^{N} ljx_ip(j,l|x_i,\Theta_t)}{\sum_{j=1}^{C}\sum_{l=1}^{\infty}\sum_{i=1}^{N} l^2j^2p(j,l|x_i,\Theta_t)} \\
\sigma_t^2 &= \frac{\sum_{j=1}^{C}\sum_{l=1}^{\infty}\sum_{i=1}^{N}(x_i-cl\mu_t)^2p(j,l|x_i,\Theta_t)}{\sum_{j=1}^{C}\sum_{l=1}^{\infty}\sum_{i=1}^{N} p(j,l|x_i,\Theta_t)}
\end{aligned}
$$

References

[1] Reality Mining Project, Human Dynamics Group at the MIT Media Lab. http://reality.media.mit.edu.

[2] UCSD wireless topology discovery project. http://sysnet.ucsd.edu/wtd/.

[3] F. Bai, N. Sadagopan, B. Krishnamachari, and A. Helmy. Modeling Path Duration Distributions in MANETs and their Impact on Routing Performance. In *IEEE Journal on Selected Areas of Communications*, September 2004.

[4] A. Balasubramanian, B. N. Levine, and A. Venkataramani. DTN Routing as a Resource Allocation Problem. In *ACM Conference on Communications Architectures, Protocols and Applications (SIGCOMM)*, 2007.

[5] J. Bilmes. A Gentle Tutorial on the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. Technical Report ICSI-TR-97-021, University of Berkeley, 1997.

[6] J. Burgess, B. Gallagher, D. Jensen, and B. N. Levine. MaxProp: Routing for Vehicle-Based Disruption-Tolerant Networks. In *IEEE International Conference on Computer Communications (INFOCOM)*, 2006.

[7] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott. Impact of Human Mobility on the Design of Opportunitic Forwarding Algorithms. In *IEEE International Conference on Computer Communications (INFOCOM)*, 2006.

[8] L.-J. Chen, Y.-C. Chen, T. Sun, P. Sreedevi, K.-T. Chen, C.-H. Yu, and H.-H. Chu. Finding Self-Similarities in Opportunistic People Networks. In *IEEE International Conference on Computer Communications (INFOCOM) Mini-Symposium*, 2007.

[9] V. Conan, J. Leguay, and T. Friedman. Heterogeneous Inter-contact Times: their Importance for DTN Routing, 2006. preprint.

[10] M. Dunbabin, P. Corke, I. Vailescu, and D. Rus. Data Muling over Underwater Wireless Sensor Networks using an Autonomous Underwater vehicle. In *International Conference on Robotics and Automation (ICRA)*. IEEE, May 2006.

[11] N. Eagle and A. Pentland. Reality Mining: Sensing Complex Social Systems. In *Journal of Personal and Ubiquitous Computing*, 2005.

[12] R. Groenevelt, P. Nain, and G. Koole. The Message Delay in Mobile Ad Hoc Networks. In *Performance*, October 2005.

[13] T. Henderson, D. Kotz, and I. Abyzov. The Changing Usage of a Mature Campus-wide Wireless Network. In *Proc. MobiCom*, 2004.

[14] W.-J. Hsu and A. Helmy. IMPACT: Investigation of Mobile-user Patterns Across University Campus using WLAN Trace Analysis. Technical report, University of South California, 2005.

[15] W.-J. Hsu, T. Spyropoulos, K. Psounis, and A. Helmy. Modeling Time-variant User Mobility in Wireless Mobile Networks. In *IEEE International Conference on Computer Communications (INFOCOM)*, 2007.

[16] W.J. Hsu and Ahmed Helmy. Encounter-based message broadcasting in ad hoc networks with intermittent connectivity. In *ACM International Symposium on Mobile Ad Hoc Networking and Computing (MOBIHOC), poster session*, 2005.

[17] R. Jain, D. Lelescu, and M. Balakrishnan. Model T: An Empirical Model for User Registration Patterns in a Compus Wireless LAN. In *MobiCom*, 2005.

[18] J. Jetcheva, Y.-C. Hu, S. PalChaudhuri, A. Kumar Saha, and D. B. Johnson. Design and evaluation of a metropolitan area multitier wireless ad hoc network architecture. In *Workshop on Mobile Computing Systems and Applications*, 2003.

[19] P. Juang, H. Oki, Y. Wang, M. Martonosi, L.-S. Peh, and D. Rubenstein. Energy-efficient computing for wildlife tracking: Design tradeoffs and early experiences with zebranet. In *ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 2002.

[20] E.L. Kaplan and P. Meier. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53, Jun 1958.

[21] T. Karagiannis, J-Y. Le Boudec, and M. Vojnovic. Power Law and Exponential Decay of Inter Contact Times between Mobile Devices. In *ACM Conference on Mobile Computing and Networking (MOBICOM)*, 2007.

[22] M. Kim, D. Kotz, and S. Kim. Extracting a Mobility Model from Real User Traces. In *IEEE International Conference on Computer Communications (INFOCOM)*, 2006.

[23] Leonard Kleinrock. *Queueing Systems, Volume 1: Theory*. John Wiley & Sons, 1975.

[24] J. Krumm and E. Horvitz. The Microsoft Multiperson Location Survey. Technical Report MSR-TR-2005-13, Microsoft Research Technical Report, August 2005.

[25] A. K. Saha and D. B. Johnson. Modeling Mobility for Vehicular Ad Hoc Networks. In *ACM international workshop on Vehicular ad hoc networks (VANET), poster session*, 2004.

[26] J. Scott, R. Gass, J. Crowcroft, P. Hui, C. Diot, and A. Chaintreau. CRAWDAD data set: cambridge/haggle (v. 2006-01-31), January 2006.

[27] J. Scott, R. Gass, J. Crowcroft, P. Hui, C. Diot, and A. Chaintreau. CRAWDAD trace: cambridge/haggle imote/infocom(v. 2006-01-31), January 2006.

[28] A. Seth, D. Kroeker, M. Zaharia, S. Guo, and S. Keshav. Low-Cost Communication for Rural Internet Kiosks Using Mechanical Backhaul. In *Proc. ACM MobiCom*, 2006.

[29] J. Su, A. Chin, A. Popivanova, A. Goel, and E. de Lara. User Mobility for Opportunistic Ad-Hoc Networking. In *IEEE workshop on Mobile Computing Systems and Applications (WMSCA)*, 2004.

[30] J. Su, A. Goel, and E. d. Lara. An Empirical Evaluation of the Student-Net Delay Tolerant Network. In *International Conference on Mobile and Ubiquitous Systems: Networks and Services (MOBIQUITOUS)*, 2006.

[31] C. Tuduce and T. Gross. A Mobility Model Based on WLAN Traces and its Validation. In *IEEE International Conference on Computer Communications (INFOCOM)*, 2005.

[32] A. Vahdat and D. Becker. Epidemic Routing for Partially Connected Ad Hoc Networks. Technical Report CS-200006, Duke University, April 2000.

[33] X. Zhang, G. Neglia, J. Kurose, and D. Towsley. Performance Modeling of Epidemic Routing. *Elsevier Computer Networks journal*, 51/10:2859–2891, 2007.