New Approaches for Characterizing Inter-contact Times in Opportunistic Networks[☆]

Enrique Hernández-Orallo*, Juan Carlos Cano*, Carlos T. Calafate*, Pietro Manzoni*

Departamento de Informática de Sistemas y Computadores. Universitat Politècnica de València. Spain.

Abstract

Characterizing the contacts between nodes is of utmost importance when evaluating mobile opportunistic networks. The most common characterization of inter-contact times is based on the study of the aggregate distribution of contacts between individual pairs of nodes, assuming an homogenous network, where contact patterns between nodes are similar. The problem with this aggregate distribution is that it is not always representative of the individual pair distributions, especially in the short term and when the number of nodes in the network is high. Thus, deriving results from this characterization can lead to inaccurate performance evaluation results.

In this paper, we propose new approaches to characterize the inter-contact times distribution having a higher representativeness and, thus, increasing the accuracy of the derived performance results. Furthermore, these new characterizations require only a moderate number of contacts in order to be representative, thereby allowing to perform a temporal modelization of traffic traces. This a key issue for increasing accuracy, since real-traces can have a high variability in terms of contact patterns along time. The experiments show that the new characterizations, compared with the established one, are more precise, even using short time contact traces.

Keywords: Opportunistic networks, Performance Evaluation, Contact-based Messaging, Inter-contact times

1. Introduction

In Opportunistic networks, contacts are sporadic and can appear intermittently so routes are built dynamically. Any contact between nodes can opportunistically be used for message relaying, provided it is likely to bring the message closer to the final destination, thus depending on cooperation to work properly. Applications of such networks include Mobile Ad-Hoc Networks (MANETs), Vehicular Ad-Hoc Networks (VANETs) and Mobile Social Networks.

Evaluating the performance of these opportunistic networks is a challenging issue. A common approach is to simulate these networks using a network simulation tool under realistic mobility traces. Nevertheless, simulation can be very time consuming and restricted to the limited scenarios

[☆] This work was partially supported by *Ministerio de Economia y Competitividad*, Spain (Grant TEC2014-52690-R) and *Generalitat Valenciana*, Spain (Grant AICO/2015/113)

[*] Corresponding author.

Email addresses: ehernandez@disca.upv.es (Enrique Hernández-Orallo), jucano@disca.upv.es (Juan Carlos Cano), calafate@disca.upv.es (Carlos T. Calafate), pmanzoni@disca.upv.es (Pietro Manzoni) of the available mobility traces. In order to avoid these drawbacks analytical models can provide a fast and broader performance evaluation. Analytical models require anyway a precise and concise description of the mobility scenario. For example, there are many analytical performance models that assume that the inter-contact times distribution between pairs of nodes are exponentially distributed with a given rate λ. For example, using a contact rate λ we can obtain the transmission delay and cost of mobile protocols, such as epidemic routing protocols [1, 2], and the impact of node selfishness on mobile networks [3, 4, 5]. The precision of the previous models clearly depends on how accurate is the estimation of the contact rate λ, which at the same time directly depends on the representativeness of the characterized distribution.

Therefore, characterizing inter-contact times (or inter-meeting times) between nodes is essential for analyzing the performance of contact based protocols such as cooperative or opportunistic networks. The established approach is to characterize the inter-contact times distribution between pairs of nodes using an *aggregate distribution* [6, 7, 8, 9, 10]. This distribution is obtained by *aggregating* the *individual pair distribution* of all node pair combinations in the network. The *individual pair distribution* is defined as the distribution of the time elapsed between two consecutive contacts between the same pair of nodes [11]. Another way to characterize inter-contact times is to consider the time elapsed between contacts for any pair of nodes in a group (known as *inter-any-contact times*). This characterization was briefly studied in [6] using human mobility traces. The conclusions were that inter-any-contact times are longer that individual pairs inter-contact times (as expected), but with a similar distribution shape. This paper also shows a time dependence in the contact distribution, with different pattern distributions for the diurnal and night periods.

Previous works have studied the distribution of the inter-contact times by collecting data from real mobile network environments [1, 7, 8, 12, 9, 13, 14]. Some of these works have shown that the aggregate inter-contact times distribution is exponential with rate λ for both human and vehicle mobility scenarios [1, 9, 13]. The work in [12] analyzed some popular mobility traces and found that over 85% of the individual pair distributions fit an exponential distribution. Nevertheless, there is some controversy about whether this exponential distribution relates to real mobility patterns. Some empirical results have shown that the aggregate inter-contact times distribution follows a power-law distribution and has a long tail [7], meaning that some pairs of nodes barely experience any contact. In [15] it is shown that in a bounded domain, the inter-contact distribution is exponential, but in an unbounded domain the distribution is power-law. The dichotomy of this distribution is described in [8], which shows a truncated power law with exponential decay appearing in its tail after some cutoff point. A recent paper [11], presented the dependence between the *individual pair distributions* and the *aggregate distribution.* It is stated that, starting from exponential *individual pair distributions*, the *aggregate distribution* is distributed according to a Pareto law. It also verifies the dichotomy property of the aggregate distribution analytically.

Summing up, most of the literature is based on the aggregate distribution, assuming that it is representative of the individual pair distribution [11]. This is the case of the so called *homogeneous* opportunistic networks assumption, where all pair contact patterns are supposed to be the same. Thus, the contact rate of the aggregate distribution is similar to the individual pair distributions contact rates. Nevertheless, as shown in [11], these contact rates are only similar when the length of the contact trace is large. Furthermore, most traces exhibit a non-homogenous behavior, where pair contact patterns are different. For example, analyzing a contact trace of a University campus we can observe that the contact pattern between students can be different to the contact pattern between staff members. Thus, obtaining a representative characterization (for example λ) of these *heterogeneous* networks to be used in analytical models is a challenging issue.

A practical approach would be to obtain an equivalent contact rate, aggregating the individual contact rates, as in the homogeneous case (*the homogeneous assumption*). However, this can lead to an inaccurate characterization, as shown in [16]. The authors of this paper compared three methods for fitting the exponential distribution from traces, always using an aggregation based distribution. These methods were evaluated using a Continuous-Time Markov Chain model of the epidemic diffusion. The results showed that none of the characterization methods used was accurate enough. Therefore, an alternative approach to accurately estimate this individual pair distribution (and the contact rate) is needed, especially for heterogeneous networks.

In this paper, we propose new approaches to improve the characterization of the inter-contact times distribution presenting higher representativeness and, thus, increasing the precision of the results obtained. Using different characterizations, three inter-contact time distributions are considered: the *Aggregate Pairs* distribution, that is the *established characterization*; the *Aggregate Nodes* distribution, that is obtained as the aggregate of inter-contact distributions between one node and the rest of nodes; and the *Any Contact* distribution, that is the inter-contact distribution between any nodes. First, we study their statistical representativeness showing that, for the same trace length, the Aggregate Pairs distribution has a very low representativeness, especially when the number of nodes is high, in contrast with the others having a good representativeness. Second, we study the relation among these distributions. We prove that, if all individual pair distributions are exponentially distributed, the *Any Contact* distribution is a new exponential distribution as well. This is not true for the *aggregate* distributions, that depends on the distribution of each individual λ [11]. The previous conclusions are very important because it allows obtaining the λ value used in the analytical models in a more precise way.

Finally, we evaluate the precision of the three distributions using both synthetic and real contact traces. The precision is evaluated using a well known analytical model, namely the epidemic message diffusion, that is based on a given exponential inter-contact times distribution with rate λ obtained from the three previous characterization methods. Experimental results confirm that the established *Aggregate Pairs* distribution is under-representative and, consequently, the precision of the results obtained using the epidemic routing model is too low, specially when the number of nodes of the evaluated network is high. Instead, the results using the *Aggregate Nodes* and the *Any-Contact* distributions are much more precise, requiring significantly smaller contact traces. Furthermore, these distributions allow the evaluation of the time dependence, obtaining more accurate results, in contrast with the low representativeness (and poor precision) of the *Aggregate Pairs* distribution.

The rest of the paper is organized as follows. In Section 2 we introduce three methods for characterizing inter-contact times distributions, evaluating their representativeness. Section 3 studies the relation between these distributions, and the associated contact rate. The experimental evaluation of the precision of the different characterizations is described in Section 4 using both synthetic and real contact traces. Finally, Section 5 presents some concluding remarks.

2. Characterizing inter-contact times distributions

In this section we describe three possible methods for characterizing the inter-contact times distribution from a contacts trace. Beside the established *Aggregate Pair* characterization we introduce two new approaches: the *Aggregate Nodes* and the *Any Contact* characterizations. We

t	n_1,n_2	d	t	n_1,n_2	d
1	(1,2)	2	7	(1,5)	3
2	(2,3)	3	8	(3,4)	5
2	(3,4)	2	11	(1,4)	4
5	(1,3)	5	14	(2,3)	2
5	(1,2)	1	17	(1,2)	3
7	(2,3)	4	20	(1,3)	6
7	(1,2)	6	20	(2,5)	2

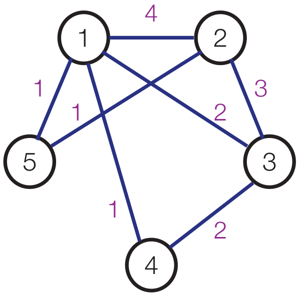


Figure 1: Contact trace sample and its contact graph. Note that a contact (n_1,n_2) means that both nodes have visibility of each other.

study their representativeness by introducing two metrics; the average number of measures and the inter-contact rate. We then evaluate the proposed metrics through some real traces.

2.1. Contact Mobility Datasets

A *contact* is defined as an opportunity of transmission between a pair of nodes (that is, two nodes are able to communicate between them directly for a given time). From all the datasets evaluated we obtained a *contact trace*.

Given a network with N nodes, the contact trace is obtained measuring the times when contacts between pairs of nodes occur for a given time interval T. The result is a trace of length $C(T)$ (that is, the number of contacts), where each record is a 4-tuple (t_i, a_i, b_i, d_i), reflecting, that at time $t_i \leq T$, there was a contact between the pair of nodes (a_i,b_i) with a duration of d_i seconds. Based on this definition, for practical issues, there is only one record for each contact between a pair of nodes (a_i,b_i), and this contact is reciprocal (in other words, there is no another record with the (b_i,a_i) contact). A simple contact trace is shown in Fig 1. This trace corresponds to a five nodes ($N = 5$) network, which has a duration T of 20 seconds, resulting in 14 contacts ($C(T) = 14$).

In order to study the extent and degree of contacts, it is practical to represent them as a graph. We use a weighted undirected graph, $G = (V, E)$ where the vertices (V) are the nodes and the edges (E) denote that there is at least one contact between the corresponding nodes. The weight of an edge $w(e)$ is the number of different contacts between two nodes. Fig 1 shows the graph for the sample contact trace. The degree of a vertex $d_G(v)$ is the number of edges incident to this vertex. A vertex of degree 0 is an isolated vertex. Finally, the maximum number of edges in an undirected graph (that is, the number of different possible contacts) is $\mathcal{E} = \frac{1}{2}|E|(|E| - 1) = \frac{1}{2}N(N - 1)$.

Using this graph, we can obtain several metrics that reflects the extent of contacts. The first one is the number of nodes with no contact, the *isolated nodes*, that is equivalent to the number of isolated vertices. The second one is the *mean degree of contacts*, corresponding to the average number of different contacts per node, that is equivalent to the mean of the vertices degrees $\bar{d}_G$. We can normalize this degree dividing it by the network size for obtaining the *degree ratio* ($\bar{d}_G/N$). Thus, a degree ratio close to one will express that practically all network nodes have contacts between them.

In this paper we use four known experimental datasets that cover a rich range of environments, from metropolitan cities to University campus, characterized by different number of nodes and trace durations. Most of the traces were obtained from the CRAWDAD repository [17]. The main characteristics of the traces are shown in table 1.

	Cambridge	Shanghai	Milano	MIT
Type	Human	Vehicle	Human	Human
Device	iMote	GPS/GPRS	Radio	Phone
Location	Campus	City-wide	Building	City-wide
Year	2005	2007	2008	2005
Network	Bluetooth	WiFi	Radio (10m)	Bluetooth
Duration (h)	274	24	276	6946
Resolution (s)	120	60	1	300
Characteristics				
Nodes (N)	36	2288	49	97
Contacts (C)	10641	1262498	11893	87007
Isolated Nodes	0	9	5	1
Mean Degree	30.06	359.96	31.14	63.94
Degree Ratio	0.835	0.157	0.636	0.659
Representativeness				
AP R_{PI}	0.811	0.043	0.596	0.536
AP M_{PI}	16.03	0.081	9.464	18.02
AN R_{NI}	1	0.993	0.898	0.989
AN M_{NI}	563.4	340.9	435.3	1737.45
AC R_{AI}	1	1	1	1
AC M_{AI}	10640	1262497	11892	87006

Table 1: Description of contact traces and representativeness of traces

1. The *Cambridge* mobility dataset [18] is a trace of bluetooth sighting gathered from a set of undergraduate students from the University of Cambridge carrying small devices (iMotes). This trace has a duration of 274 hours (11 days) and has 36 mobile nodes (students). Although this dataset has also static nodes, we only evaluate the mobile nodes, so contacts with static nodes were removed from the trace. As shown in Table 1, this trace exhibits high degree ratios with no isolated nodes.

2. The *Shanghai* Taxis GPS dataset [9] was collected from 2100 taxis in Shanghai city during February of 2007. This trace does not contain the contacts (it contains GPS locations), so a pre-process for obtaining the contact trace is needed. Following the method used in [9] we assume that a contact occurs if both vehicles are in WiFi range (100 meters). Due to the higher number of nodes, the degree ratio of this trace is relatively low and there are a few isolated nodes.

3. The *Milano* dataset [19] is a high resolution dataset collected at the University of Milan during 15 working days in November 2008, in an area comprised by offices and laboratories. Contacts were logged by custom radio devices with a transmission range of 10 meters and 1 second resolution. Regarding the contact characteristics, the degree ratio is moderately high with five isolated nodes.

4. The *MIT* (or Reality) dataset [20] was collected from 97 MIT students and staff carrying mobiles phones for about nine months. These phones logged contact with other bluetooth devices by doing device discovery every five minutes. We processed this trace, in a way similar to [16]. First, we took into account a contact by either of the two nodes (the reason

is that the original trace sometimes reflects only a one way contact, that is, node a sees node b). Second, we merge multiple consecutive contacts if their inter-contact duration time was less than one second.

2.2. Characterizations

Using a contact trace, we can characterize the inter-contact times distribution. This distribution is influenced by the trace's duration and resolution [7]. The resolution is defined as the smallest interval between two successive measurements. Inter-contact times that last more or are close to the duration of the experiment, and inter-contact times that last less than the time resolution, cannot be observed.

The inter-contact times can be modeled as a renewal process. Let $\{X_i | i = 1, 2, \ldots\}$ be a sequence of non-negative random variables, where $X_i : i > 1$ are independent and identically distributed with finite average inter-contact time ($E[X_i] < \infty$). X_i represents the inter-contact times between contacts, that is, the time between the i^{th} contact and the $(i+1)^{th}$ contact. Finally, X_i has a distribution function $F(x) = P(X \leq x)$, called the *underlying distribution.* X_1 is a random variable, not necessarily with the same distribution of $X_i : i > 1$, denoting the time until the first contact occurs (that is, the process is delayed). The renewal sequence $\{S_n\}$ can be obtained using the following expression:

$$S_n = \sum_{i=1}^{n} X_i \tag{1}$$

so S_n is the time of the n^{th} contact. If we define $\mu = E[X_i] : i > 1$, as the average inter-contact time, we can see that:

$$E[S_n] = E[X_1] + \cdots + E[X_n] = (n-1)\mu + E[X_1] \tag{2}$$

so, as a delayed renewal process, the average inter-contact time can be obtained as:

$$\mu = \frac{E[S_n] - E[X_1]}{n-1} \qquad n > 1 \tag{3}$$

Using the renewal sequence, we define the counting function as:

$$N(t) = \max\{n : S_n \leq t\} \tag{4}$$

that provides the number of contacts up to time t. Finally, the renewal function $m(t) = E[N(t)]$ is the average number of contacts (renewals) up to time t and by the elementary renewal theorem we have:

$$\lim_{t \to \infty} \frac{m(t)}{t} = \frac{1}{E[X_i]} = \frac{1}{\mu} \tag{5}$$

This expression represents the average contact rate[1], that is, $\lambda = \mu^{-1}$.

Now, we formally describe the three characterizations, the *Aggregate Pairs* (AP), the *Aggregate Nodes* (AN) and the *Any Contact* (AC). For clarity of exposition, we are going to use the simple

[1] Abusing notation, we denote the average contact rate using λ. This does not imply that the distributions must be exponential.

contact trace of Fig 1. Thus, from a contact trace we can characterize three possible inter-contact distributions.

The *Aggregate Pairs* (AP) distribution is the aggregate of inter-contact times distributions between the same pair of nodes. This is the established characterization, usually known as simply the *inter-contact times* distribution [1, 7, 8, 12, 9]. This distribution is obtained by aggregating the *individual pair inter-contact times* of all node pair combinations in the network. We need at least two contacts between the same pair of nodes in order to obtain an inter-contact time. For example, using our contact trace, for the pair (1,2) we have the following contact times[2]: {1,5,7,17} and so the inter-contact times are {4,2,10}. We can also obtain the inter-contact times for the following pair of nodes: (1,3), (2,3) and (3,4); however, for the rest of pairs, this is not possible. By aggregating all the inter-contact times calculated for the previous pairs, we can obtain the Aggregate Pairs distribution: {4,2,10,15,5,7,6}. Thus, the Aggregate Pairs (AP) has a renewal process $\{X_i^{AP}\}$ that is the union (aggregation) of the renewal process of contacts between pairs of nodes $\{X_i^p\}$:

$$\{X_j^{AP}\} = \{ \bigcup_{p \in \mathcal{P}(T)} X_i^p \} \tag{6}$$

where $\mathcal{P}(t)$ denotes the set of pairs of nodes that have at least one contact up to time t. Note that this is a synthetic process and does not represent any real sequence of contacts. Instead, it depends on how the random variables are arranged (in our contact trace example, we simply arrange the inter-contact times concatenating the contacts between pairs). Nevertheless, this arrangement does not affect our study, as our goal is to study the underlying distribution, that is, the AP distribution $F_{AP}(x)$, that does not depend on this arrangement.

The *Aggregate Nodes* (AN) distribution is the aggregate of inter-contact distributions between one node and the rest of nodes. For example, using the sample trace, we can obtain the contacts of node 1 with the other nodes: (1,x)={1,5,5,7,7,11,17,20}, so the inter-contact times are {4,0,2,0,4,6,3}. For this trace all nodes have contacts with other nodes. By aggregating all the inter-contact times we can obtain the Aggregate Nodes distribution: {4,0,2,0,4,6,3,1 ,3,2,0,7,3,3,0,3,2,1,6,6,6,3,13}. As the AP distribution, the Aggregate Nodes has a renewal process $\{X_i^{AN}\}$ that is the union (aggregation) of the renewal process of contacts between one node and the rest of nodes $\{X_i^n\}$. That is:

$$\{X_i^{AN}\} = \{ \bigcup_{n \in \mathcal{N}(T)} X_i^n \} \qquad i > 1 \tag{7}$$

where $\mathcal{N}(t)$ is the set of nodes that have at least one contact. Finally, this process has an underlying distribution $F_{AN}(x)$.

The *Any Contact* (AC) distribution is the inter-contact distribution between all nodes. This is known as the *inter-any-contact times* in [6]. In this case there is only one distribution of inter-contact times, which corresponds to the difference between two consecutive contacts. For our sample trace we have 13 inter-contact time values: {1,0,3,0,2,0,0,1,3,3,3,3,0}. In this case, the Any

[2] In this paper, the inter-contact time is computed as the difference between the starting times of two consecutive contacts. Another way to compute this time (see [7]) is to obtain the difference between the end of a contact and the start of the next one. In this case, the duration of the contact is used to obtain the inter-contact times.

Contact (AC) has a renewal process $\{X_i^{AC}\}$ where the random variables are the inter-contact times between any contact, and its underlying distribution function is $F_{AC}(x)$.

Regarding the graph representation, note that the *Aggregate Pairs* distribution refers to the edges between two vertices, and the *Aggregate Nodes* to the edges incident to a vertex. For the AP distribution, a pair of nodes has a given inter-contact time if the corresponding vertices are connected with an edge with a label greater than one. For the AN distribution, a node has a contact if, for the corresponding vertex the sum of the degrees of the incident edges are greater than zero.

2.3. Representativeness of characterizations

A critical factor for any characterizations is its representativeness. This representativeness depends on the trace length $C(T)$, the number of nodes N in the network, and the characterization used. Two metrics are defined to measure this representativeness: the *inter-contact ratio* and the *average number of measures* per individual distribution. These metrics are detailed below for each distribution. First, we define the indicator function, $\mathbf{1}(x)$ as:

$$\mathbf{1}(x) = \begin{cases} 1 & \text{if } x \text{ is } true \\ 0 & \text{if } x \text{ is } false \end{cases} \tag{8}$$

and for the Aggregate Pairs (AP) distribution, the inter-contact ratio for aggregate pairs is the ratio of pairs that have at least two contacts:

$$R_{PI} = \frac{1}{\mathcal{E}} \sum_{p \in \mathcal{P}(T)} \mathbf{1}(C_P(p,T) > 1) \tag{9}$$

where $C_P(p,T)$ is the number of contacts between pairs of nodes $p \in \mathcal{P}(T)$ (that is, the weight of the edge that connects the vertices of the associated graph), and $\mathcal{E}$ the number of different possible contacts ($\mathcal{E} = \frac{1}{2}N(N-1)$). For our sample trace, this ratio is $R_{PI} = 4/10 = 0.4$. The average number of measures M_{PI} is computed as the sum of all inter-contact times divided by $\mathcal{E}$, that is:

$$M_{PI} = \frac{1}{\mathcal{E}} \sum_{p \in \mathcal{P}(T)} (C_P(p,T) - 1) \tag{10}$$

So, in our example, we have M_{PI}=$(3+2+1+1)/10$=0.7.

In the Aggregate Nodes (AN) distribution, the inter-contact ratio is:

$$R_{NI} = \frac{1}{N} \sum_{n \in \mathcal{N}(T)} \mathbf{1}(C_N(n,T) > 1) \tag{11}$$

where $C_N(n,T)$ is the number of contacts of node n, that is the sum of the degree of the incident edges for the corresponding vertices. For our sample trace, this ratio is $R_{NI} = 5/5 = 1$. The average number of measures M_{NI} is computed as:

$$M_{NI} = \frac{1}{N} \sum_{n \in \mathcal{N}(T)} (C_N(n,T) - 1) \tag{12}$$

So, in our example, we have M_{NI}=$(7+6+6+2+1)/5$=4.2.

Finally, in the any contact distribution (AC) the ratio is 1 (if $C(T) > 1$) and the number of measures is $C(T) - 1$.

As expected, we can see that the representativeness of the AP distribution is very low: a contact ratio of only 0.4, and less than one measure per pair of nodes. In the AN distribution the representativeness is greater: there is a full contact ratio and a mean of 4.2 measures for each node. The greatest representativeness corresponds to the AC distribution, which has a full ratio and the highest number of measures.

In general, for the same trace time T, the lowest representativeness is for the AP distribution, and this representativeness (the values of the inter-contact ratio and average measures) is inversely proportional to $\mathcal{E}$, that exhibits quadratic growth with N. Therefore, when the number of nodes is high, the representativeness is very low. For the AN distribution, its representativeness is inversely proportional to the number of nodes N. Overall, the AC distribution provides the best representativeness.

We study the representativeness of the previous distributions using the real contact traces and the results are shown in Table 1. Regarding the representativity metrics for the AP distribution, when focusing on the Cambridge trace, which has only 36 nodes, the contact ratio distribution is high (although not one), and the average number of measurements is relatively high. We can say that it has a moderate representativeness. The Milano trace exhibits a lower contact-ratio but a higher number of measures per pair. For the Shanghai trace, the one with a high number of nodes, the AP distribution has a very low representativeness, so this characterization is not useful for obtaining information about the contact pattern on this network. These results show that, when the number of nodes is relatively high, the representativeness of the AP distribution is very low. For the AN distribution, all traces have a good representativeness (that is, the inter-contact ratio is close to one and the average measures are high). And finally, for the AC distribution, all traces have excellent representativeness (full ratio and the highest number of measures).

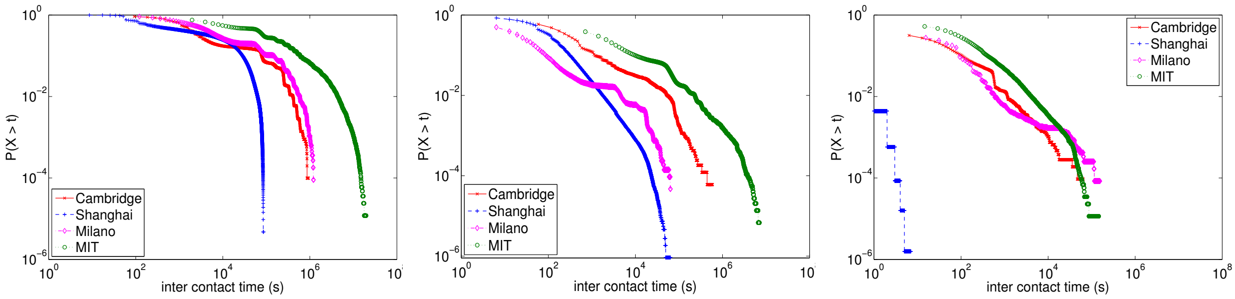
In the next subsection we are going to evaluate the different distributions using the datasets and the different methods for distribution fitting.

2.4. Evaluation of distributions

In order to study the underlying distribution $F(x)$ of inter-contact times, the CCDF (Complementary Cumulative Distribution Function) is widely used [7, 8, 11]. The CCDF (also known as the *tail distribution*) is useful to study how often a random variable is above a particular level, that is, $\bar{F}(x) = P(X_i > x)$.

Fig 2 shows the CCDF of the different distributions for the four traces evaluated. In the CCDF for the AP characterization, we can clearly observe the dichotomy of the distribution: the inter-contact time follows a power-law decay up to a characteristic time (about $10^6 s \approx 270$ hours for the MIT set, $10^5 s \approx 12$ hours for the Cambridge and Milano datasets, and $10^4 s \approx 3$ hours for Shanghai dataset) and beyond this time, the decay is exponential. This confirms earlier studies [8, 9]. For the AN distribution, the previous dichotomy is not so evident, and the exponential decay is clearer. Finally, for the AC distribution, we can see an exponential decay for the Cambridge and MIT traces. For the Shanghai trace there are resolution issues: the inter-contact times are below the resolution time, and so its values are very discretized.

Now, we are going to study the distribution fitting of the inter-contact times. The common approach is to assume that all random variables are identically exponentially distributed, that is, the *homogeneous assumption.* This implies, that all nodes in the network have as similar mobility and contact pattern. Thus, we can model this contact pattern using a single average contact



(a)

(b)

(c)

Figure 2: CCDF of the different characterizations for the dataset. a) AP distribution, b) AN distribution, c) AC distribution

rate parameter (λ) that characterizes the exponential distribution. A simple exponential fit is to estimate the average contact rate as the inverse of the average of inter-contact times of the traces, as detailed in [12]:

$$\lambda = \frac{C(T)}{\sum_{i=1}^{C(T)} s_i} \tag{13}$$

where s_i are the inter-contact time samples obtained using the different characterizations. Based on this fitting method, and using always the AP characterization, the values obtained for λ are: $1.25 \times 10^{-4}s^{-1}$ (0.45 contacts/hour) for the Milano trace [16] and $2.8 \times 10^{-5}s^{-1}$ (0.101 contacts/hour) for the Cambridge trace [13].

The problem with this characterization based on the AP distribution is it low representativity, leading to non significant results. One of the main problem is that many pairs of nodes never meet (for example, in the Milano are 40%) so we do not have information about their distribution. We can compensate these missing contacts by assuming that, in these cases, they meet with some average inter-contact time T' that is always greater or equal than the trace duration T, so we can obtain a new average inter-contact time. This approach was used in [16], and for $T' = 276h$, the contact rate λ for the Milano trace is $1.51 \times 10^{-5}s^{-1}$ (0.054 contacts/hour).

Another approach is based on not taking into account the higher inter-contact times samples, as the fit performed in [9] for the Shanghai trace. The reason is that inter-contact times that last longer than the duration of the trace data cannot be observed and those with very large values close to the duration are less likely to be found. Consequently, the weight of large inter-contact times in the distribution is biased. Thus, their approach consists on identifying a divide point on the CCDF graph, and perform an exponential fit for values less than this point, obtaining a value $\lambda = 3.71 \times 10^{-6}s^{-1}$ (0.013 contacts/hour) for the Shanghai trace.

A similar approach is based on log-normal fitting. Some distributions seems to have a better fit with a log-normal distribution [14]. Thus, the authors in [16] proposed to apply curve fitting with log-normal distribution and then estimate the average inter-contact times. Based on this approach, the contact rate for the Milano dataset is $\lambda = 6.73 \times 10^{-5}s^{-1}$ (0.24 contacts/hour), that is a rate that is between the simple exponential fit and the one obtained by compensating the missing contacts. Nevertheless, the experiments described in this paper indicate that this is not a good estimation.

The heterogeneous network model is used in [21, 22, 13]. In this model, the heterogeneity is characterized by allowing different contact rates between node pairs with rates λ_p. Note that this model can be extremely cumbersome, as the number of pairs (that is, rates) increases quadratically with the number of nodes. A practical approach is to group similar contact patterns under different social groups (*social heterogeneity*) [21, 23] or spatial cluster sites (*spatial heterogeneity*) [24], where each group has a different contact rate.

Summing up, in this section we have introduced two new approaches to characterize the inter-contact distribution: the *Aggregate Nodes* (AN) and the *Any Contact* (AC). According to the representativity metrics, these characterizations have a greater representativity than the established approach, that has a very reduced representativity when the number of nodes is high. These facts were confirmed using real datasets with different characteristics. Regarding the exponential distribution fitting, there are several approaches (all based on the *Aggregate Pairs* characterization), but none of them seems a good estimation.

In the next section, we are going to study the relation between the distributions of the previous characterizations, in order to obtain a better estimation of the exponential distribution, and thus, of the corresponding λ values.

3. Relation between distributions

In this section we study the relation between the different distribution characterizations (AP, AN, AC). Our main goal is to establish relations between the individual distributions and the aggregate distributions. The established approach is to obtain the individual pair contact rate (λ_p) from the aggregate pairs inter-contact rate (λ_{AP}). This relation is of special interest, as many analytical models based on Markov Chains use λ_p as the contact rate. In this section we introduce new approaches to obtain λ_p from the AN and AC distributions that are more representative. As we will show, if the individual pair distributions are not exponential we cannot obtain a simple relation between these distributions. Finally, we also study the relation between the average contact rates that are valid for any distribution.

3.1. Relation between individual and aggregated distributions

In this subsection we study the relation between the aggregated distributions (AP and AN) and the individual distributions (between pairs and nodes). In general, this relation is not simple, specially for heterogeneous networks (that is, when individual distributions are different).

According to the notation introduced in subsection 2.2, the Aggregate Pairs (AP) renewal process $\{X_i^{AP}\}$ is the union of the renewal process of contacts between pair of nodes. Let us arrange the sequence of random variables as follows:

$$X_1^{AP} = X_1^1, X_2^{AP} = X_2^1, \dots, X_{c_1}^{AP} = X_{c_1}^1, X_{c_1+1}^{AP} = X_1^2, \dots, X_{c_1+c_2}^{AP} = X_{c_2}^2, \dots \tag{14}$$

where c_p is the number of contacts of pair p up to time T, that is, $C_P(p,T)$. The renewal sequence $\{S_n^{AP}\}$ for this arrange is the accumulated time of the i^{th} contact of the p^{th} pair. We can obtain the time of the last contact $n = C(T) = \sum_{p\in\mathcal{P}(T)} c_p$ as[3]:

$$\begin{aligned} E[S_n^{AP}] =& E[X_1^1] + \cdots + E[X_{c_1}^1] + E[X_1^2] + \cdots + E[X_{c_2}^2] + \cdots \\ =& (c_1 - 1)\mu_1 + (c_2 - 1)\mu_2 + \cdots + E[X_1^1] + E[X_1^2] + \cdots \\ =& \sum_{p\in\mathcal{P}(T)} (c_p - 1)\mu_p + \sum_{p\in\mathcal{P}(T)} E[X_1^p] \end{aligned} \tag{15}$$

where μ_p is the average inter-contact time of X_i^p. We can observe this process as a delayed process where the first $|\mathcal{P}(T)|$ values of $\{S_n^{AP}\}$ correspond to delay. Then, applying an equation equivalent to expression 3, we have:

$$\mu_{AP} = \frac{\sum_{p\in\mathcal{P}(T)}(c_p - 1)\mu_p}{n - |\mathcal{P}(T)|} = \sum_{p\in\mathcal{P}(T)} \frac{C_P(p,T) - 1}{C(T) - |\mathcal{P}(T)|}\mu_p \tag{16}$$

so we can see that the average inter-contact time of the aggregated process is the weighted average of the individual average inter-contact times.

The relation between the underlying distributions $F_{AP}(x)$ and $F_p(x)$ of the previous renewal processes was established in [8] through the CCDF as:

$$\bar{F}_{AP}(x) = \sum_{p\in\mathcal{P}(T)} \frac{C_P(p,T) - 1}{C(T) - |\mathcal{P}(T)|}\bar{F}_p(x) \tag{17}$$

that is similar to expression 16. When T tends to be large,

$$\bar{F}_{AP}(x) = \sum_{p\in\mathcal{P}} \frac{\lambda_p}{\Sigma\lambda_p}\bar{F}_p(x) \tag{18}$$

where $1/\lambda_p$ is the average of inter-contact times for the pair p (in other words, λ_p is the average contact rate of the pair), and $\Sigma\lambda_p = \sum_{p\in\mathcal{P}} \lambda_p$. Expression 18 shows that the aggregate CCDF is equal to the weighted sum of individual CCDF with a weight proportional to the rate of contacts λ_p. Furthermore, if all individual CCDF ($\bar{F}_p(x)$) are identical, and if the ratio of contacts (R_C) is assumed to be 1 (note that if $T \to \infty$ then $R_C \to 1$), then the aggregate CCDF and the individual CCDF are the same. As a result, if all individual distributions are exponentially distributed with λ_p, then the aggregate distribution is exponentially distributed with $\lambda_{AP} = \lambda_p$. Based on equation 18, when the individual distributions are different and exponentially distributed, we cannot obtain a simple relation between these distributions and the aggregate one. More information is needed about the distribution of the individual rates λ_p and, depending on this distribution, the aggregate distribution can follow an Exponential, Pareto, or Power Law with Exponential Decay distribution (see paper [11] for a detailed study about these relations).

In a similar way, we can establish the following relation for the Aggregate Nodes (AN) distribution:

$$\mu_{AN} = \sum_{n\in\mathcal{N}(T)} \frac{C_N(n,T) - 1}{C(T) - |\mathcal{N}(T)|}\mu_n \tag{19}$$

[3] Note that for $n = C(T)$, the value of $E[S_n^{AP}]$ does not depend on how the random variables are arranged, so the following expression is true in any case.

and the relation between distributions $F_{AN}(x)$ and $F_n(x)$ through the CCDF is:

$$\bar{F}_{AN}(x) = \sum_{n\in\mathcal{N}(T)} \frac{C_N(n,T)-1}{C(T)-|\mathcal{N}(T)|}\bar{F}_n(x) \tag{20}$$

and when T tends to be large:

$$\bar{F}_{AN}(x) = \sum_{n\in\mathcal{N}} \frac{\lambda_n}{\Sigma\lambda_n}\bar{F}_n(x) \tag{21}$$

where $1/\lambda_n$ is the mean value for the inter-contact times for node n, and $\Sigma\lambda_n = \sum_{n\in\mathcal{N}}\lambda_n$. If all individual CCDF ($\bar{F}_n(x)$) are identical and $R_C = 1$, then the aggregate CCDF is equal to the individual CCDF, and the exponential distributions have the same contact rate ($\lambda_n = \lambda_{AN}$).

Summing up, in this subsection we have established the relation between individual and aggregate distributions. Except for the case when all individual distributions are equally distributed, this relation is not simple. In the following subsection we will show that the relation between the Any contact distribution and the individual distribution is simpler, and this allow obtaining the λ_p value of the individual exponential distribution from any of the distributions.

3.2. Relation between individual and Any Contact distributions

For the Any Contact (AC) renewal process $\{X_i^{AC}\}$, we can see that if a contact occurs at time t, the next contact time will be the minimum time of all possible contacts between pairs, subsequent to time t. Thus, using the renewal sequence, we have:

$$\{S_n^{AC}\} = \{\min_{p\in\mathcal{P}(T)}\{S_{n_p}^p\}\} \qquad \forall S_{n_p}^p \geq S_n^{AC} \tag{22}$$

This minimum depends on the type of distribution of the associated renewal process. Fortunately, for Poisson processes this expression has a simple solution. For n independent exponentially distributed random variables with rate parameters $\lambda_1, \ldots \lambda_n$, the minimum is also exponentially distributed with parameter $\lambda = \lambda_1 + \cdots + \lambda_n$. Then, if all random variables X_i^p are exponentially distributed, the X_i^{AC} is also exponentially distributed:

$$F_{AC}(x;\lambda_{AC}) = F(x;\sum_{p\in\mathcal{P}(T)}\lambda_p) \tag{23}$$

where λ_{AC} is the contact rate of the AC distribution. We can see that this relation is simpler than expressions 18, and 21 and we can obtain the average contact rate of the exponential AC distribution as:

$$\lambda_{AC} = \sum_{p\in\mathcal{P}}\lambda_p \tag{24}$$

Furthermore, assuming that all individual pair random variables X_i^p are identically distributed it allows estimating the contact rate λ_{AC}, when T tends to be large as:

$$\lambda_{AC} = |\mathcal{P}|\lambda_p = \frac{1}{2}N(N-1)\lambda_p = \frac{1}{2}N(N-1)\lambda_{AP} \tag{25}$$

Note that when T tends to be large, the number of different posible contacts $|\mathcal{P}|$ tends to its maximum $\mathcal{E} = \frac{1}{2}N(N-1)$.

The relation between the AC distribution and the AN distribution is obtained in a similar way, through the renewal sequence:

$$\{S_m^{AC}\} = \{\min_{n\in\mathcal{N}(T)} \{S_{m_n}^n\}\} \qquad \forall S_{m_n}^n \geq S_m^{AC} \tag{26}$$

In this case, we can see that for a given pair (a, b), the random variable is repeated twice, one for the first node $a \in p$, and another one for the second node $b \in p$. For the AC distribution the contacts are no longer repeated, and if all distributions are exponentially distributed, we have:

$$F_{AC}(x; \lambda_{AC}) = F(x; \sum_{n\in\mathcal{N}(T)} \lambda_n) \tag{27}$$

Assuming that all individual pair random variable X^n are identically distributed we have:

$$\lambda_{AC} = \frac{1}{2} N\lambda_n = \frac{1}{2} N\lambda_{AN} \tag{28}$$

when T tends to be large. Combining expressions 25 and 28 we have that:

$$\lambda_p = \lambda_{AP} = \frac{\lambda_{AN}}{N-1} = \frac{\lambda_{AC}}{\frac{1}{2}N(N-1)} \tag{29}$$

so λ_p can be estimated using any of the previous distributions. This relation is very important, as we can directly estimate the AP contact rate, that is usually adopted in analytical models by using the more representative AN and AC distributions.

In conclusion, we have shown that the AN distribution have a simpler relation with the individual pair distribution than the AP distribution. Furthermore, if the individual pair distributions are exponentially distributed, the AN distributions is also exponentially distributed. This is also true for the individual node distribution. This fact confirms the exponential shapes shown in Fig 2.

3.3. Relation between Contact Rates

In this section we study the relation between the contact rates of the different characterizations, based on the individual contact rates. The following relations are valid for all types of distributions (not only for exponential distributions). We will see that these relations are equivalent to the ones obtained in the previous subsection.

The relation between the average contact rate of the AC distribution (λ_{AC}) and the average contact rates of individual pair of nodes (λ_p) is obtained as follows. For the Any Contact renewal process $\{X_i^{AC}\}$, the average number of contacts generated up to time T can be obtained through the renewal function $m_{AC}(T)$. For large values of T, $m_{AC}(T) = T \cdot \lambda_{AC}$. The average number of contacts for the individual pair renewal processes $\{X_i^p\}$ is $m_p(T) = T \cdot \lambda_p$. Since the number of contacts must be the same for the $\{X_i^{AC}\}$ process and the sum of all the individual processes $\{X_i^p\}$, we have $m_{AC}(T) = \sum_{p\in\mathcal{P}(T)} m_p(T)$. Finally, removing T on both parts, we have:

$$\lambda_{AC} = \sum_{p\in\mathcal{P}(T)} \lambda_p \tag{30}$$

The relation between λ_{AC} and the individual node contact rate λ_n is obtained in a similar way. For the $\{X_i^n\}$ renewal process, we have that the average number of contacts is $m_n(T)$. In this case, for obtaining the sum of all contacts for all nodes, we can see that, for a given pair (a, b), a contact is repeated twice, one for the first node on the process $\{X_i^a\}$, and another one for the second node on the process $\{X_i^b\}$. Thus, the sum of contacts, $\sum_{n\in\mathcal{N}(T)} m_n(T)$, is twice the number of contacts for the AC process, where the contacts are no longer repeated, and so we have:

$$\lambda_{AC} = \frac{1}{2} \sum_{n\in\mathcal{N}(T)} \lambda_n \tag{31}$$

Now, we can obtain the relation between the average contact rate of the AP distribution (λ_{AP}) and λ_p. First, for the renewal process $\{X_i^{AP}\}$, we can obtain the expected time for the last contact using expression 15 as the sum of the expected times of all individual process:

$$E[S_n^{AP}] = E[S_n^1] + E[S_n^2] + \ldots = \sum_{p\in\mathcal{P}(T)} E[S_n^p] \tag{32}$$

When T tends to be large, we find that the expected time for the last contact $n = C(T)$, $E[S_n^{AP}]$ is $|\mathcal{P}(T)|T$. Then $m_{AP}(|\mathcal{P}(T)|T) = |\mathcal{P}(T)|T \cdot \lambda_{AP}$ is the total number of contacts. This number of contacts must be equal to the sum of contacts of all renewal processes $\{X_i^p\}$ up to time T, that is $\sum_{p\in\mathcal{P}(T)} m_p(T) = T \cdot \sum_{p\in\mathcal{P}(T)} \lambda_p$. Thus, making equal both expressions and simplifying, we have:

$$\lambda_{AP} = \frac{\sum_{p\in\mathcal{P}(T)} \lambda_p}{|\mathcal{P}(T)|} \tag{33}$$

Finally, the relation between the average contact rate for the AN distribution λ_{AN} and λ_n is obtained in a similar way:

$$\lambda_{AN} = \frac{\sum_{n\in\mathcal{N}(T)} \lambda_n}{|\mathcal{N}(T)|} \tag{34}$$

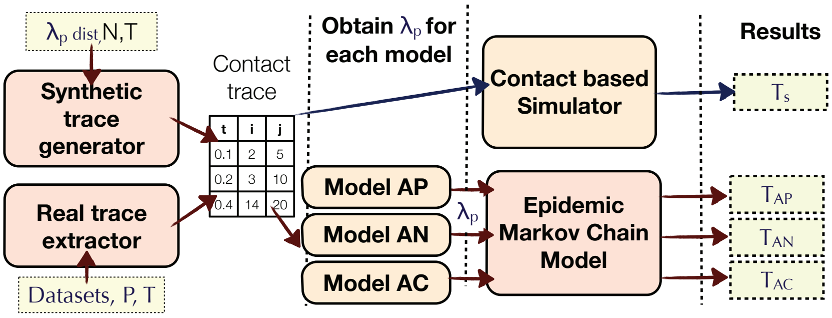
That is, the contact rate of the aggregate distribution is equal to the mean of the individual contact rates. Finally, using expressions 33 and 34, we can obtain a relation between contact rates that is equivalent to expression 29.

4. Experimental Evaluation

In the previous sections we have studied the representativeness of the different distributions and the relation between them. We showed that the average node (AN) distribution, and especially the any contact (AC) distribution, have better statistical representativeness than the widely established Aggregate Pairs (AP) distribution. Now, we are going to evaluate the precision of these characterizations.

There are many analytical performance models that assume that the inter-contact times distribution between pairs of nodes is exponentially distributed [1, 2, 3, 9]. This fact allows using the contact rate λ in Markovian models, such as the epidemic diffusion model. These models assume a unique contact rate λ between all pair of nodes. So, all individual pair distributions must be equal and exponentially distributed with mean λ_p. As shown in section 3, when assuming the previous condition (the *homogeneous assumption*), and for a T large enough, λ_p can be estimated from any of the previous distributions using expression 29.

Thus, to evaluate the precision, we proceed by comparing the results of using an analytical model with the different λ obtained using expression 29 with the results obtained using the contact



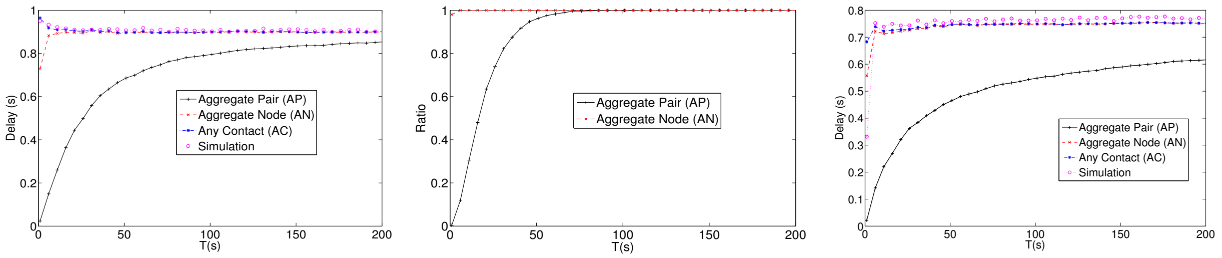
t	i	j
0.1	2	5
0.2	3	10
0.4	14	20

Figure 3: The evaluation process. Note that the contact trace can be generated synthetically or from the real dataset.

trace in a simulator. Concretely, we use the model for the Epidemic Routing. The model we use, based on Markov chain models, was introduced in [1] for obtaining the average source-to-destination delivery delay ($\tau = E[T_d]$). Thus, we compare the delivery delays using this model with the results obtained using a simulator. A closer result will reflect a more precise characterization. This approach was also used in [16]. Additionally, we also evaluated the precision using other performance models (such as the Two Hop multi-copy protocol, and a more sophisticated model for evaluating the selfish detection introduced by the authors in [5]), obtaining similar results in terms of precision than the ones described in this paper.

The process for evaluating the distributions is depicted in Fig 3. We synthetically generate a contact trace for a given time T with inter-contact times between pairs of nodes that are exponentially distributed with a contact rate λ_p. This contact trace is used to estimate three different contact rates $\hat{\lambda}_p$ from the different contact rates of the distributions (λ_{AP},λ_{AN},λ_{AC}) using expression 29. Then, using the different $\hat{\lambda}_p$ generated for each distribution (AP, AN, AC) we obtain the delivery time using the Markov model for epidemic routing. For validation purposes, we also implemented a custom simulator. This simulator reads the contact trace and simulates the transmission of a message between a randomly selected pair of nodes in the network. The simulation finishes when the message reaches the destination node obtaining the delivery time. This evaluation is repeated 1000 times (that is, 1000 different traces are generated) in order to obtain a mean value. We also obtain the ratio of contacts on the simulator, that is similar to the ratio of contacts (R_C) obtained for evaluating the representativeness of the distributions. In case a simulation ends without the message being delivered to the destination node, we count this as a miss, and when the experiments end, we obtain the delivery ratio as $(1000 - misses)/1000$.

In the real trace experiments the process is very similar, except that in this case we used a real traffic trace as the input of the model and simulator. In order to evaluate the precision of the different distributions depending on the duration of the traffic trace, the original traffic trace is trimmed from T ranging from 0 to the maximum duration (that is, only the contacts with time $t \leq T$ are used). In the periodic evaluation, we select a period P, and for each time interval i, we extract the contacts that are in the range $[Pi, P(i+1)[$. This contacts trace is then used to obtain the delay time using the evaluation process depicted in Fig 3. Finally, although the traffic trace is always the same, the simulation is repeated 1000 times, varying the pair of selected sender and destination nodes of the network.



(a)

(b)

(c)

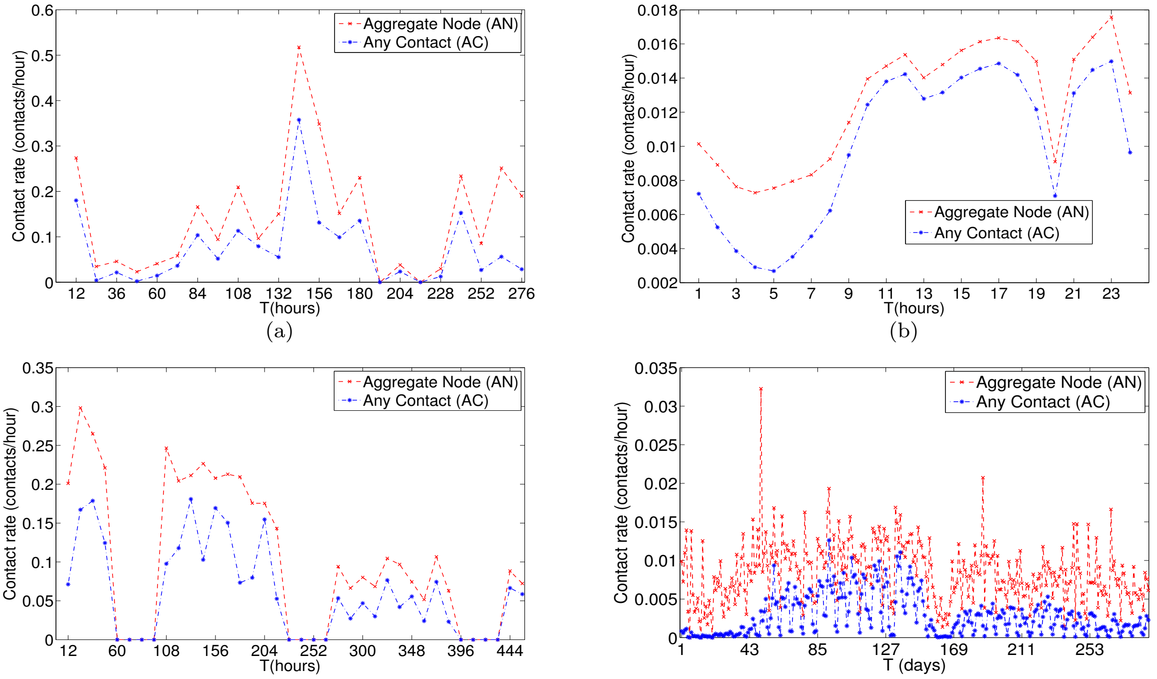
Figure 4: Evaluation of the precision using synthetic traces. a) Delay of a message using an epidemic routing transmission with an homogeneous trace, b) Contact ratio of the homogeneous trace (R_C), c) Delay with an heterogeneous trace

4.1. Synthetic traces

The following experiments used synthetic traces to validate the conclusions derived about representativeness and precision in sections 2 and 3. Fig 4a shows the delivery delay for the different distributions calculated using a synthetic trace with time T ranging from 1 to 200s that was generated assuming that all inter-contact times distributions between pairs are exponentially distributed with the same contact rate $\lambda_p = 0.1$ (that is, all distributions are homogeneous) and for a network of 50 nodes. We can see that the AN and AC distributions obtain the best results, converging very fast to the simulation results (for example, for $T = 200s$, the delay obtained through simulation is 0.918s and the delays obtained using the AN and AC distributions are the same: 0.905s, a relative error of 1.42%). Instead, the AP distribution converges to the simulation results very slowly. As expected, these results clearly depend on the representativeness of the different distributions as shown in Fig 4b. The AP contact ratio is very low for a reduced trace time (T), reaching 1 for $T = 100$. Nevertheless, for this trace time, the average number of measures is still very reduced ($M_{PI} = 9.6$ for $T = 100$ in the AP distribution compared to $M_{NI} = 765.86$ in the AN distribution), so the precision is also low. Thus, as expected, the conclusions are clear, the most precise characterizations are the *Aggregate Nodes* (AN) and the *Any Contact* (AC).

In the following experiment we study the precision of the *homogeneous assumption* in the case of heterogeneous exponential distributions. We generate a synthetic contact trace with the contact times between pairs of nodes exponentially distributed, where the contact rate λ_p is also exponentially distributed with mean 0.1. From the generated trace, we estimate an equivalent single contact rate, that is used on the models to obtain the delays of the epidemic transmission. These values are compared to the simulated one as shown in Fig 4c. In this case we can see that the delay obtained using the AC and AN distribution is very precise (a relative error of 2.34% for $T = 200s$). Nevertheless, the delay obtained for the AP distribution converges to the solution very low. The reason is that, for an heterogeneous distribution the representativeness of the AP distribution is lower than in the homogeneous case (for example, for $T = 200s$, $R_{PI} = 0.92$ and $M_{PI} = 23$). The rate of contact reaches 1 for approximately $T = 2000s$ with $M_{PI} = 233$, and a relative error of 7.24%.

From these synthetic evaluations we can estimate how large T must be so that expression 29 can be applicable, i.e. the results are representative and accurate. This effect depends mainly on the number of contacts generated up to time T, that is $C(T)$. In the previous experiments, we can estimate this number of contacts as $C(T) = T(N-1)N\lambda_p$. Using this expression, for the



(a) (b) (c) (d)

Figure 5: Average Contact rate depending on time (contacts/hour). a) Cambridge dataset (period = 12 hours); b) MAWI dataset (period = 1 hour); c) Milano dataset (period = 12 hours) d) MIT data set (period = 1 day).

	Cambridge	Shanghai	Milano	MIT
AP	0.1352	0.0454	0.0836	0.0127
AN	0.0535	0.0094	0.0282	0.0029
AC	0.0616	0.0101	0.0330	0.0027

Table 2: Average contact rate λ_p obtained for the different traces. All the values are in contacts/hour.

AP distribution, we obtain representative results for $T = 100s$, that is, $C(100) = 24500$ contacts. Instead, for the AN distribution, the value of T for representative results is approximately 5 seconds, so the number of contacts is 1225. Finally, the AC distribution needs about 200 contacts.

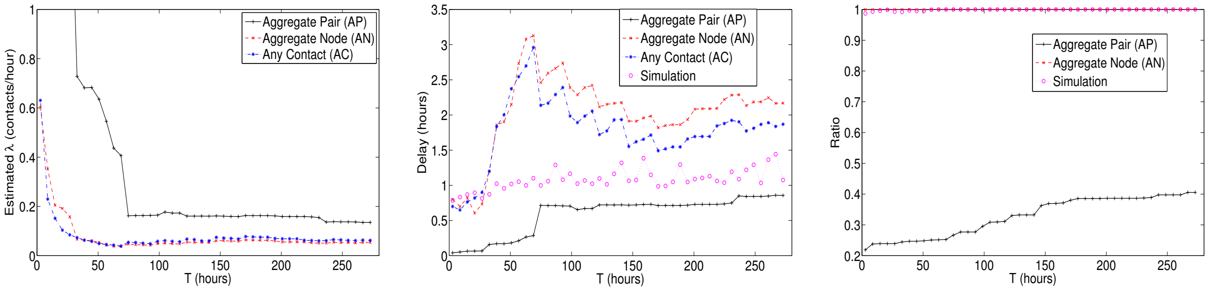
Summarizing, we can see that for the AN and AC a reduced number of contacts in needed in order to obtain accurate results, thus allowing the evaluation of traces using smaller time intervals. Additionally, AN and AC are more accurate characterizations than AP in the case of heterogeneous networks.

4.2. Real traces

We now evaluate the precision of the different distributions using real traffic traces. We estimated the different values of $\hat{\lambda}_p$ for the traces evaluated, as shown in Table 2. These values where obtained using the whole trace. In general, we can see that the $\hat{\lambda}_p$ obtained using either AN and AC characterizations are similar. However these values differ from the ones obtained using the AP characterization. This is especially evident for the Shanghai trace. The main reason is the lower representativity of the AP characterization.

	Cambridge	Shanghai	Milano	MIT
AP	0.857	0.008	1.093	4.178
AN	2.169	0.388	3.237	17.841
AC	1.884	0.361	4.099	19.764
SIM	1.412	2.623	94.237	1195.382

Table 3: Delivery time for the epidemic protocol using the model with the three different λ estimations and the value obtained using the simulation. All the value are in hours.



(a) (b) (c)

Figure 6: Evaluation of the precision using the Cambridge traces depending on trace duration T. a) Estimated contact rate ($contacts/hour$), b) Delay of a message using an epidemic transmission, c) Contact ratio (R_C).

Using the previous contact rates we proceed to obtain the delivery delay using the epidemic model and compare it against the simulated values. The results are shown in Table 3, which shows some differences with respect to the simulated ones. This is especially evident for the Milano and MIT traces. These results confirm the experiments conducted by [16], that also showed that a single homogeneous contact rate is unable to capture the global performance of a contact trace.

The main reason for this behavior is that contact rates are very variable between different periods on the same day (for one day traces), and also between different days (for traces that last several days), as we can see in Fig 5. For example, Fig 5a shows the different average contact rates for the Cambridge trace when selecting a period of $P = 12$ hours. Using this period, we estimate the different values of $\hat{\lambda}_p$ for the different distributions by normalizing them using expression 29. The rates obtained using either the AN and AC distributions are very similar, but the rates obtained using the AP distribution are quite different, and so their are not presented due to its low representativeness. For the Shanghai trace using a period of 1 hour we can clearly observe in Fig 5b that the contact rate is low during the night, starting to increase after 8 am. In the Milano dataset we can easily determine the weekend, where contact rates are zero. Finally, the MIT trace shows a huge variation of contact rates between working and weekends days, where the contact rate is almost zero.

Thus, a way to increase the precision of the estimated results is to reduce the duration of the selected period. For shorter periods (for example hours), the contact rate is expected to have less variability. Fig 6 shows the impact of this period on the results for the Cambridge trace depending on a trimmed traffic trace with duration T. First, we can see in Fig 6a how the estimated value of λ decreases with the trace duration T. Thus, using this contact rate, we can obtain the delay, as shown in Fig 6b. We can see that the results obtained using the AC and AN distributions are close to the simulation results. There is a strong time dependence of this trace (the sawtooth shape),

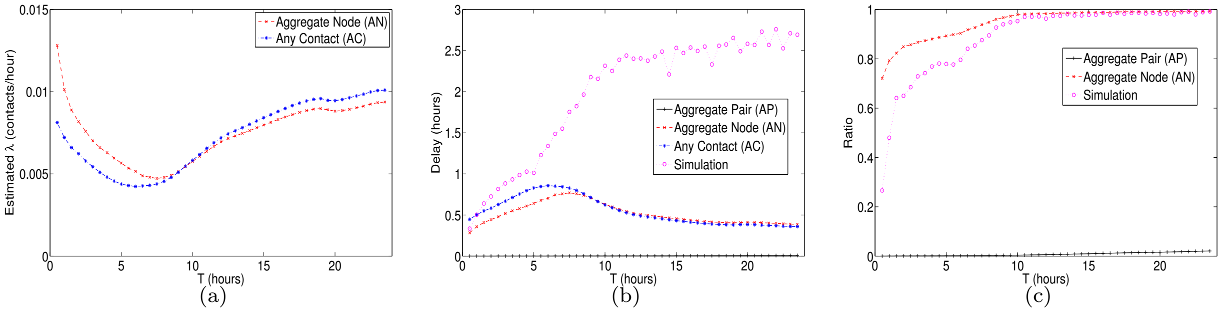


Figure 7: Evaluation of the precision using the Shanghai traces. a) Estimated contact rate (*contacts/hour*), b) Delay of a message using an epidemic transmission, c) Contact ratio (R_C).

due to the high differences between contact rates in diurnal and night periods. Regarding the AP distribution, the precision increases as the contact ratio increases (see Fig 6c), but it is still far away from those results obtained using the other distributions.

Fig 7 shows the results for the Shanghai traffic traces. The estimated contact rate is variable for the AN and AC distributions. The results for the AP distribution are not shown, because they are several orders of magnitude higher. This is due to the low representativeness of this distribution (see Fig 7c). Fig 7b shows the delay obtained using the AN and AC distributions for trimmed traffic traces for T values less than 20000 (about 5 hours) are close to the simulated ones, although the results for greater values of T show a higher error. We can also observe how the contact ratio of the simulation is low for values less than 10000s, so the representativeness of the simulation is also low.

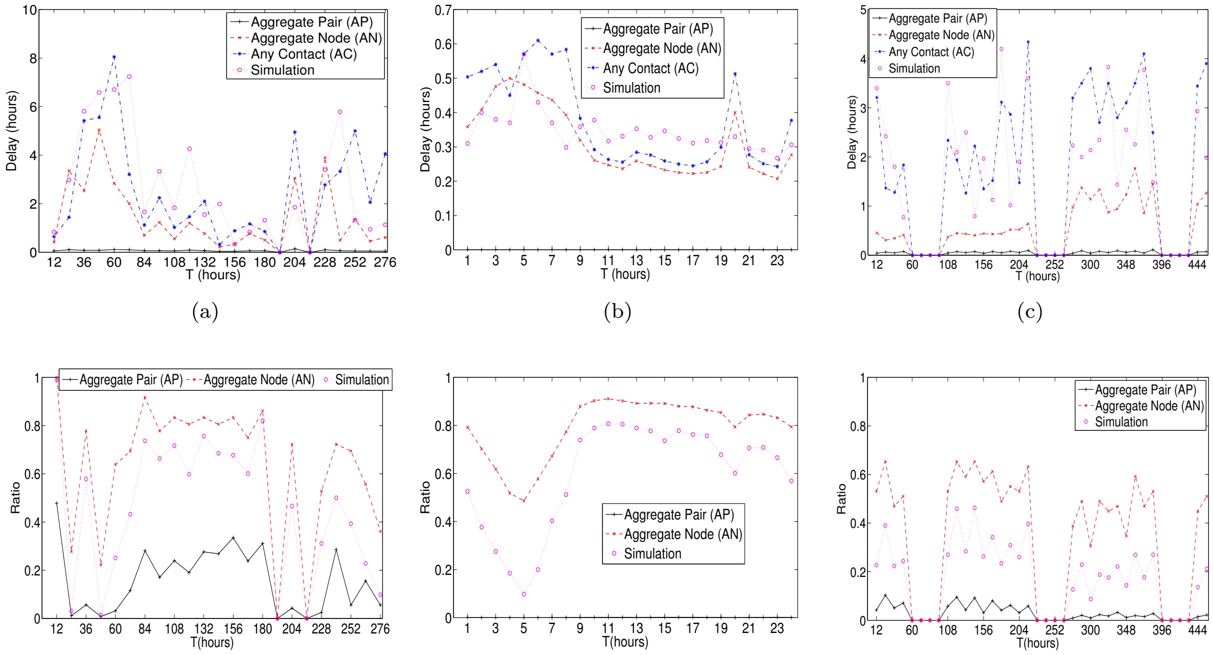
The conclusions drawn from the previous experiments for the AN and AC distributions are clear: the greater the contact trace duration, the lesser the precision. This is mainly due to the long-tail behavior of the inter-contact times distribution between the same pair of nodes and the temporal variability of the contact rates. If the duration of the trace is high (days in the case of Cambridge, or hours in the case of Shanghai), the probability of contact between some reduced number of node pairs is very low, so the inter-contact time becomes very high. In other words, the greater the duration of the trace, the greater the tail of the distribution[4].

4.3. Periodic evaluation

As shown in Fig 5, we can observe from real-traffic traces that there is a great variation on the contact rate between different periods on the same day, or between different days. Thus, using a single contact rate for the whole trace cannot reflect this variation. It will be more precise to evaluate a given period, obtaining a given contact rate. For example, when studying the Shanghai contact rate, we can make a different evaluation for hours with low traffic and hours with higher traffic.

Now we proceed to perform a periodic evaluation of the traces. We estimate, for each period, the average contact rates λ (using the three characterizations), and compare the obtained delays using the epidemic model with the simulated ones. The period P for each dataset is the same used in Fig 5. For the Cambridge dataset (see Fig 8a) the results obtained for the AN and AC

[4] Note that we are studying the average behavior. Thus, if we make a good estimate of the tail distribution, it can lead to wrong average results.



(a)

(b)

(c)

(d)

(e)

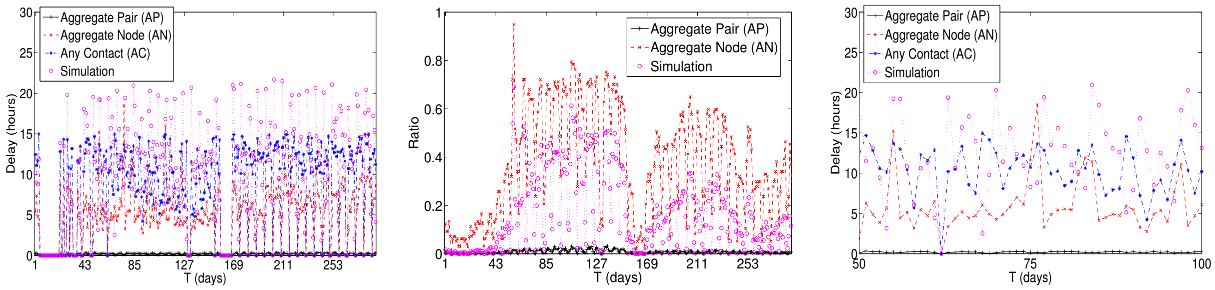
(f)

Figure 8: Time dependence of traces (left column plots are for Cambridge trace, center column plots are for Shanghai trace and right column for Milano a-c) Packet transmission delay, and d-f) Contact ratio (R_C)

distributions are close to the simulated ones, following the same pattern. As expected, the results for the AP distribution significantly differ from the simulated ones due to their low representativeness, as shown in Fig 8d. For the Shanghai dataset, the obtained results using the AN and AC distributions are precise (Fig 8b), while for the AP distribution the error is high due to the very low representativeness of this distribution.

Fig 8c shows the delay for the Milano dataset. The results for AC are still close to the simulation results. The AN results are not so close due to their moderate representativity (see Fig 8f). In order to evaluate the precision of the previous results we should consider the moderate representativity of the simulation results due to the reduced number of contacts in that trace. Finally, we evaluated the precision using the MIT traces. In this case, we discard the results when the period includes less than 100 contacts (in one day) because the simulation results are not representative. Fig 9a shows that the results are less accurate. This is due to its low resolution and high variability. Fig 9c shows the previous results for a shorter time interval (50 days). In general, we can see that the delay obtained using the AN and AP distributions is similar to the simulated values, and that the results using AP are several order of magnitude lower than the simulated ones.

We can evaluate the global accuracy of each characterization using the *order of magnitude error* (OME) or *mean logarithmic error*. For a given period P, and for the different characterizations (AP, AN, AC), we obtained a set of diffusion delays τ_i^m using the epidemic diffusion model. These values are compared with the simulated ones τ_i^s. Thus, for a trace of length T, we have K values



(a)

(b)

(c)

Figure 9: Time dependence of for the MIT traces a) Packet transmission delay, b) Contact ratio (R_C), c) Packet transmission delay for 50 days.

Trace (Period)	AP	AN	AC
Cambridge (12 hours)	1.595	0.234	0.212
Cambridge (6 hours)	1.371	0.251	0.325
Shanghai (2 hours)	3.157	0.292	0.224
Shanghai (1 hour)	2.998	0.208	0.282
Milano (24 hours)	1.749	0.651	0.309
Milano (12 hours)	1.560	0.512	0.358
MIT (24 hours)	1.737	0.360	0.417
MIT (12 hours)	1.665	0.354	0.583

Table 4: Order of magnitude error for the different dataset and periods.

($K = \lceil T/P \rceil$). We define order of magnitude error as:

$$OME = \frac{1}{K} \sum_{i=1}^{K} |log_{10}\tau_i^m - log_{10}\tau_i^s| \tag{35}$$

A value less than one reflects that the values obtained using the model are in same order of magnitude that the simulated ones. Table 4 shows the OME for the different dataset and different periods. As shown on previous graphs, the results for the AP characterization show an error several orders of magnitude higher than AN and AC.

Summing up, the previous experiments show that, when using short time trace durations, the obtained results are in general accurate for the AN and AC distributions. This is not true for the AP distribution, due to its lower representativeness when the trace duration T is low. Finally, regarding which distribution (AN or AC) to choose, we found that in general both characterizations provide similar results.

5. Conclusions

In this paper we introduced new approaches to characterize the inter-contact times distribution in order to increase the representativeness, and thus the precision of the analytical performance models based on these exponential distributions. The usefulness of exponential characterizations and models is based on its simplicity. Using only two parameters (contact rate and number of nodes) we can model a high variety of protocols, such as epidemic diffusion, two-hop, selfish node detection, among others.

We have seen that the established characterization, the *Aggregate Pairs* (AP) distribution, has very low representativeness especially when the number of nodes is high and the number of contacts is low. This leads to poor results when applied to analytical models (such as the epidemic diffusion). Therefore, we introduced and evaluated two alternative methods for characterizing the inter-contact times distribution: the *Aggregate Nodes* (AN) and the *Any-Contact* (AC). The resulting distributions have an excellent representativeness for short trace durations. We also obtained a simple relation between the individual contact rate and the contact rate for the AP and AN distributions. Thus, we can use these distributions in order to obtain the individual pair contact rate used in analytical models.

The experiments confirm that the representativeness and precision achieved using these new distributions are higher. Using synthetic and real traces we showed that the *Aggregate Nodes* (AN) and *Any-Contact* (AC) distributions are more precise than the *Aggregate Pairs* (AP) distribution. Furthermore, using the AN and AC distributions allows to partition the trace for making a periodic evaluation obtaining results that are several order of magnitude more precise that the ones obtained using the AP distribution.

References

[1] R. Groenevelt, P. Nain, G. Koole, The message delay in mobile ad hoc networks, Performance Evaluation 62 (2005) 210–228.

[2] X. Zhang, G. Neglia, J. Kurose, D. Towsley, Performance modeling of epidemic routing, Computer Networks 51 (10) (2007) 2867 – 2891.

[3] M. Karaliopoulos, Assessing the vulnerability of DTN data relaying schemes to node selfishness, Communications Letters, IEEE 13 (12) (2009) 923 –925.

[4] E. Hernández-Orallo, M. D. Serrat, J.-C. Cano, C. M. T. Calafate, P. Manzoni, Improving selfish node detection in MANETs using a collaborative watchdog, IEEE Communications Letters 16 (5) (2012) 642–645.

[5] E. Hernández-Orallo, M. D. Serrat Olmos, J.-C. Cano, C. T. Calafate, P. Manzoni, Evaluation of collaborative selfish node detection in MANETS and DTNs, in: Proceedings of the 15th ACM international conference on Modeling, analysis and simulation of wireless and mobile systems, MSWiM '12, ACM, New York, NY, USA, 2012, pp. 159–166.

[6] P. Hui, A. Chaintreau, J. Scott, R. Gass, J. Crowcroft, C. Diot, Pocket switched networks and human mobility in conference environments, in: Proceedings of the 2005 ACM SIGCOMM workshop on Delay-tolerant networking, WDTN '05, ACM, New York, NY, USA, 2005, pp. 244–251.

[7] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, J. Scott, Impact of human mobility on opportunistic forwarding algorithms, IEEE Transactions on Mobile Computing 6 (2007) 606–620.

[8] T. Karagiannis, J.-Y. Le Boudec, M. Vojnović, 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, MobiCom '07, ACM, New York, NY, USA, 2007, pp. 183–194.

[9] H. Zhu, L. Fu, G. Xue, Y. Zhu, M. Li, L. M. Ni, Recognizing exponential inter-contact time in VANETs, in: Proceedings of the 29th conference on Information communications, INFOCOM'10, IEEE Press, Piscataway, NJ, USA, 2010, pp. 101–105.

[10] A. Passarella, M. Conti, C. Boldrini, R. I. Dunbar, Modelling inter-contact times in social pervasive networks, in: Proceedings of the 14th ACM international conference on Modeling, analysis and simulation of wireless and mobile systems, MSWiM '11, ACM, New York, NY, USA, 2011, pp. 333–340.

[11] A. Passarella, M. Conti, Characterising aggregate inter-contact times in heterogeneous opportunistic networks, in: Proceedings of the 10th international IFIP TC 6 conference on Networking - Volume Part II, NETWORKING'11, Springer-Verlag, Berlin, Heidelberg, 2011, pp. 301–313.

[12] W. Gao, Q. Li, B. Zhao, G. Cao, Multicasting in delay tolerant networks: a social network perspective, in: Proceedings of the tenth ACM international symposium on Mobile ad hoc networking and computing, MobiHoc '09, ACM, New York, NY, USA, 2009, pp. 299–308.

[13] Y. Li, G. Su, D. Wu, D. Jin, L. Su, L. Zeng, The impact of node selfishness on multicasting in delay tolerant networks, Vehicular Technology, IEEE Transactions on 60 (5) (2011) 2224 –2238.

[14] V. Conan, J. Leguay, T. Friedman, Characterizing pairwise inter-contact patterns in delay tolerant networks, in: Proceedings of the 1st international conference on Autonomic computing and communication systems, Autonomics '07, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), ICST, Brussels, Belgium, Belgium, 2007, pp. 19:1–19:9.

[15] H. Cai, D. Y. Eun, Crossing over the bounded domain: From exponential to power-law intermeeting time in mobile ad hoc networks, Networking, IEEE/ACM Transactions on 17 (5) (2009) 1578 –1591.

[16] L. Pajevic, G. Karlsson, O. Helgason, Epidemic content distribution: empirical and analytic performance, in: Proceedings of the 16th ACM international conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems, MSWiM '13, ACM, New York, NY, USA, 2013, pp. 335–340.

[17] U. of Dartmouth, CRAWDAD data set, Downloaded from http://crawdad.cs.dartmouth.edu.

[18] J. Leguay, A. Lindgren, J. Scott, T. Friedman, J. Crowcroft, Opportunistic content distribution in an urban setting, in: Proceedings of the 2006 SIGCOMM Workshop on Challenged Networks, CHANTS '06, ACM, New York, NY, USA, 2006, pp. 205–212.

[19] S. Gaito, E. Pagani, G. Rossi, Fine-grained tracking of human mobility in dense scenarios, in: Proceedings of the 6th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks Workshops, 2009. SECON Workshops '09., 2009, pp. 1–3.

[20] N. Eagle, A. Pentland, Social serendipity: mobilizing social software, Pervasive Computing, IEEE 4 (2) (2005) 28–34.

[21] V. Conan, J. Leguay, T. Friedman, Fixed point opportunistic routing in delay tolerant networks, Selected Areas in Communications, IEEE Journal on 26 (5) (2008) 773–782.

[22] C.-H. Lee, D. Y. Eunt, Heterogeneity in contact dynamics: Helpful or harmful to forwarding algorithms in dtns?, in: Proceedings of the 7th International Conference on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks, WiOPT'09, IEEE Press, Piscataway, NJ, USA, 2009, pp. 72–81.

[23] T. Spyropoulos, T. Turletti, K. Obraczka, Routing in delay-tolerant networks comprising heterogeneous node populations, Mobile Computing, IEEE Transactions on 8 (8) (2009) 1132–1147.

[24] N. Banerjee, M. D. Corner, D. Towsley, B. N. Levine, Relays, base stations, and meshes: Enhancing mobile networks with infrastructure, in: Proceedings of the 14th ACM International Conference on Mobile Computing and Networking, MobiCom '08, ACM, New York, NY, USA, 2008, pp. 81–91.