# Epidemic Content Distribution: Empirical and Analytic Performance

Ljubica Pajevic, Gunnar Karlsson, Ólafur Helgason
KTH Royal Institute of Technology
Linnaeus ACCESS Center
Stockholm, Sweden
{ljubica, gk, olafurr}@kth.se

## ABSTRACT

Epidemic content dissemination has been proposed as an approach to mitigate frequent link disruptions and support content-centric information dissemination in opportunistic networks. Stochastic modeling is a common method to evaluate performance of epidemic dissemination schemes. The models introduce assumptions which, on one hand make them analytically tractable, while on the other, ignore attested characteristics of human mobility. In this paper, we investigate the fitness and limitations of an analytical stochastic model for content dissemination by comparison with experimental results obtained from real mobility traces. Our finding is that a homogeneous analytic model is unable to capture the performance of content dissemination with respect to content delivery delays.

### Categories and Subject Descriptors

C.2.1 [**Network Architecture and Design**]: Wireless comunication; C.4 [**Performance of Systems**]: Modeling techniques

### Keywords

Epidemic modeling, opportunistic networks, ad hoc networks, content distribution

## 1. INTRODUCTION

Opportunistic networking is seen as a feasible way to provide communication between mobile devices in absence of infrastructure, or as a mean to off-load existing mobile networks. Both cases are seen as scalable solutions for content distribution that adopts a *store-carry-and-forward* paradigm: a node downloads and stores contents; it then carries the contents through its mobility and shares the contents with other nodes it encounters.

A variety of opportunistic routing schemes have been proposed and epidemic spreading is central to many. Epidemic content distribution schemes are able to achieve minimum delivery delay at the expense of increased use of resources, such as buffer space, transmission power, and bandwidth. In order to exploit trade-off between delivery delay and resource consumption, different schemes limit the number of hops for contents to be carried. Adopting the principles of epidemic modeling from the field of mathematical biology to study spreading of diseases, stochastic modeling has become a common approach.

In this paper, we empirically study the performance of epidemic content spreading by using real-world mobility traces. Then, we consider an analytic model proposed in [8], and examine if this homogeneous model can be utilized to evaluate the performance of opportunistic networks. We consider a basic epidemic scheme, where all nodes participate in content forwarding.

The rest of the paper is organized as follows. In Section 2 we describe the analytic model for epidemic content distribution. We present the empirical study and compare it with analytic evaluation in Section 3. Section 4 summarizes related work and in Section 5 we draw main conclusions of this study and give directions for our future work.

## 2. OPPORTUNISTIC CONTENT DISTRIBUTION MODEL

### 2.1 Application scenario

The application scenario we consider here is that of disseminating information by utilizing opportunistic contacts and based on user interest. Sharing local news, traffic and tourist information in public areas, public announcements at massive events, or mobile advertisements are common examples where this can be used. From the perspective of a publisher, questions of interest could be: ”how many users will the information reach in a period of one hour?”, or ”what is the probability that the information will reach a certain number of users?”. Such questions can be answered by analytic modeling of information spreading.

### 2.2 Homogeneous system model

We consider a network $\mathcal{N}$ with $|\mathcal{N}| = N$ mobile nodes, equipped with short-range radios and moving in a bounded area. The network is assumed to be relatively sparse, with node density insufficient to establish a connected network. The data is stored and carried by nodes, and transferred through intermittent contacts occurring owing to node mobility. Let us introduce the definitions and assumptions that we will use in this text. The *contact time* is the duration of time when two nodes are in transmission range of each other. The *inter-contact time* for a pair of nodes is defined as the time elapsed between two consecutive contacts.

*MSWiM'13,* November 3–8, 2013, Barcelona, Spain.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2353-6/13/11
http://dx.doi.org/10.1145/2507924.2507998

We assume that the mobility of nodes is such that the inter-contact times between any pair of nodes can be modelled by independent identically distributed (i.i.d.) random variables that are exponentially distributed. Then, we assume that nodes in the network are *homogeneous*, that is, all the nodes have the same mobility and contact patterns that follows the same exponential inter-contact distribution with average rate $\lambda$.

The content spreading scheme works as follows. At time $t = 0$, there is a single node in the network that possesses the content item and all other nodes in the network are interested in obtaining it. Nodes that obtained the content are willing to forward the content to other nodes they meet. We study the performance by investigating the time it takes for the content to reach all the other $N - 1$ nodes. The transmissions are assumed to be instantaneous and every contact results in a successful transmission.

We are interested in two metrics which characterize the process of content distribution, namely *overall* and *individual delivery time*. Consider a network $\mathcal{N}$ and an arbitrarily chosen node $i$, $i \in \mathcal{N}$. Given that content is available at $i$ at time $t = 0$, the *overall delivery time*, denoted by $T_{odt}$, is the time until the content has reached all the other $N-1$ nodes. The time until a node $j$, $j \in \mathcal{N}$ has obtained the content is the *individual delivery time*, $T_{idt}$. From a performance perspective, $T_{odt}$ measures the performance of the entire system, while $T_{idt}$ is a measure of the system performance seen from an arbitrary node.

Borrowing the terms from epidemic modeling [3], we denote nodes that carry contents as *infected*, and nodes that are interested in obtaining contents as *susceptible* nodes. In our model, once a susceptible node is infected, it stays in that state for the remainder of the epidemic process.

## 2.3 Stochastic model

In the field of epidemic modeling, there are two main approaches to analyse spreading: stochastic and fluid-based modeling. Stochastic models are preferable when studying networks of small scale, as they allow some randomness in the final number of infected. The fluid models present a deterministic approximation of the stochastic spreading and can therefore only produce accurate results for networks of larger scale. From an engineering point of view, only stochastic models are able to predict the distribution of time until a certain percentage of network has been infected. Herein, we consider a stochastic model for content distribution, based on a continuous-time Markov chain. We present only the basic description referring interested readers to [8] for the complete analysis.

Let the random variable $X(t)$ be the number of infected nodes at time $t, t \geq 0$ with $X(0) = 1$. Since all the $i$ infected nodes spread the content further, the process $\{X(t); t \geq 0\}$ is a pure-birth process with with rates $\lambda_i = i(N-i)\lambda$ for all the states $i = 1, ..., N-1$.

$T_{odt}$ is the time it takes the system to reach the absorbing state $X(T_{odt}) = N$. The time the system spends in each transient state $i$ is exponentially distributed with the expected value $1/\lambda_i$, and the average absorption time is given by the sum:

$$E[T_{odt}] = \sum_{i=1}^{N-1} \frac{1}{\lambda_i} = \frac{1}{\lambda} \sum_{i=1}^{N-1} \frac{1}{i(N-i)} = \frac{2}{\lambda N} H_{N-1} \qquad (1)$$

where $H_n = \sum_{i=1}^{n} 1/i$ is the $n$-th harmonic number.

To obtain $E[T_{idt}]$, denote by the random variable $T_{k,N-1}$ time until $k$ out of the $N-1$ susceptible nodes have become infected, and introduce the event $K$ that a given node is the $k$-th to become infected. Since all nodes are identical and inter-contact times are i.i.d., the probability of this event is $Pr\{K\} = 1/(N-1)$ and

$$E[T_{idt}] = \sum_{k=1}^{N-1} E[T_{k,N-1}] Pr\{K\} = \frac{1}{N-1} \sum_{k=1}^{N-1} E[T_{k,N-1}].$$

Omitting a few steps of derivation (complete proof in [8]), the expected individual delivery time is

$$E[T_{idt}] = \frac{1}{\lambda(N-1)} H_{N-1}. \qquad (2)$$

Expressions (1) and (2) describe epidemic spreading with respect to contents delivery times. In the next section, we will empirically analyse the spreading in four scenarios, by means of simulation.

# 3. ANALYSIS WITH MOBILITY TRACES

## 3.1 Mobility Datasets

To cover various scenarios, we use four experimental datasets different in time granularity, number of participants and in duration. The datasets report pairwise contacts between users moving in relatively restricted areas: a conference venue, a university and office buildings. Note, however, that the areas are not strictly bounded, thus users may leave the areas and return after longer periods (even days). Below we describe the contexts where the traces were collected, the acquisition methodologies used, and our methodology of pre-processing the traces.

**Infocom** mobility traces [12] were obtained during four days at Infocom 2006. The dataset reports direct contacts between a group of 78 attendees of a workshop, who were carrying iMotes. The scanning interval was 120 seconds. Thus, it is likely that many shorter contacts were not recorded. Also, contacts between two users scanning simultaneously are missing. Therefore, we assume that two nodes were in contact if either of the nodes reported that event. As our model assumes a closed system (no nodes leaving), we consider only the time intervals when most of the nodes were active and seen by other users. The experiment lasted for four days; we extracted from the trace only the contacts during daytime: between 9:00 and 18:00 on the first and the third day, between 9:00 and 21:00 on the second, and between 9:00 and 16:00 on the fourth day.

**Humanet** trace [1] describes human mobility of participants in an office building. The data collection was carried out in a company building and contains traces of 52 participants, company employees, during one working day. The users were carrying Bluetooth devices, which were scanning

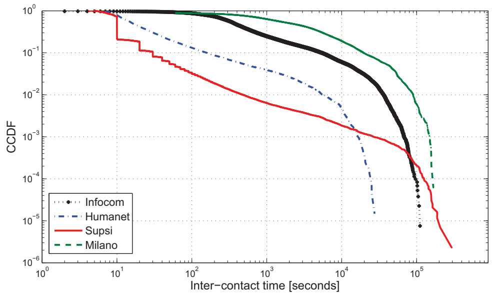
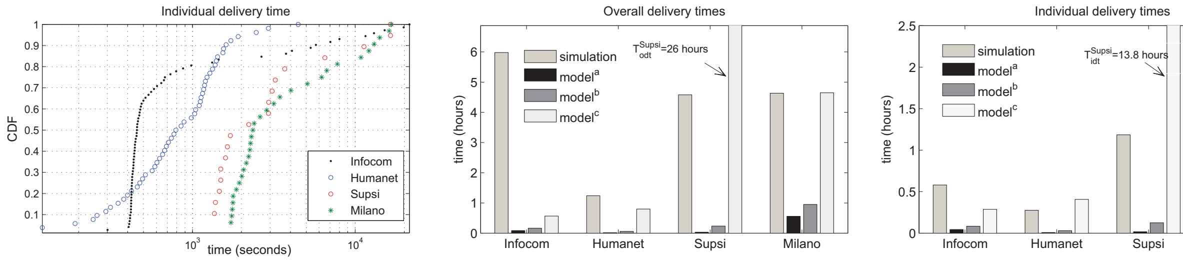


**Figure 1: CCDF of aggregate inter-contact times.**



(a)

(b)

(c)

**Figure 2: (a) CDFs of the delivery times. Comparison of the (b) overall and (c) individual delivery times.**

every 5 seconds to capture direct contacts with other devices. For each user, contact entries contain the time when the contact started and when it ended. First, we processed the trace to account for all the contacts recorded by either of the two nodes, and when both nodes recorded the same event, we took the entry with longer contact duration. Some entries contained contacts of zero duration; if either of two nodes reported a contact with non-zero duration, we chose that entry. The next step was to merge multiple consecutive contacts of zero duration if their inter-contact time was shorter than one second into a single contact with the duration equal to the sum of their inter-contact times, and finally, omit contacts that occurred before 10:00 or after 19:00.

**Supsi** dataset [4] includes contacts between 39 participants from three institutes, located in two buildings. The experiment was carried out in December 2010 and lasted more than three weeks. We use records of eleven days when the largest number of contacts was recorded. Proximity information was collected by sensor nodes, carried by the users. The nodes were configured to have a transmission range of 5 meters and perform neighbour discovery every 10 milliseconds. Similarly as with the previous traces, we consider only contacts from 9:00 to 18:00. This comes from the assumption that the contacts took place in the area of interest, and that they represent real mobility of people, as some devices collected contact information during the night when left by the users in the offices. This leaves records of 34 users in total; the number of active users per day varied from 13 to 27.

**Milano** dataset [5] was collected at the University of Milano in November 2008 from 44 mobile devices carried by faculty members, graduate students, and technical staff. The experiment area comprised offices and laboratories located in a three-floor building, and nearby premises where participants took breaks during lunch times. Contacts were logged by devices operating with a transmission range of 10 meters and a configurable scanning interval of around one second. By using the same procedure of filtering out sparse contacts during the night or during the days when few participants were active, we extract only the contacts during work hours from 9:00 to 18:00 over twelve days.

Note that all traces capture direct contacts between the experiment participants, and, aside from the *Infocom* trace, with scanning intervals in the order of seconds. Due to the long scanning interval, the *Infocom* trace may be missing shorter contacts.

## 3.2 Aggregate inter-contact time distributions

We calculate the inter-contact times between any two nodes, and assume that all the samples come from one and the same distribution, the *aggregate* inter-contact time distribution. The distribution of samples of inter-contact times for a specific pair of nodes is denoted by *pair-wise* inter-contact time distribution. The aggregate distributions are plotted in Fig. 1. The average inter-contact times of these distributions are: 2347 (*Infocom*), 312 (*Humanet*), 323 (*Supsi*) and 7996 seconds (*Milano*) ($\bar{\tau}_{ag}$ in Tab. 1). Only the *Milano* trace seems to resemble exponential distribution, and only up to around 6 hours (found by looking carefully into a lin-log scale). All distributions exhibit fast decay after a certain value (usually order of hours), which however, could simply be an artefact of the finite duration of the traces.

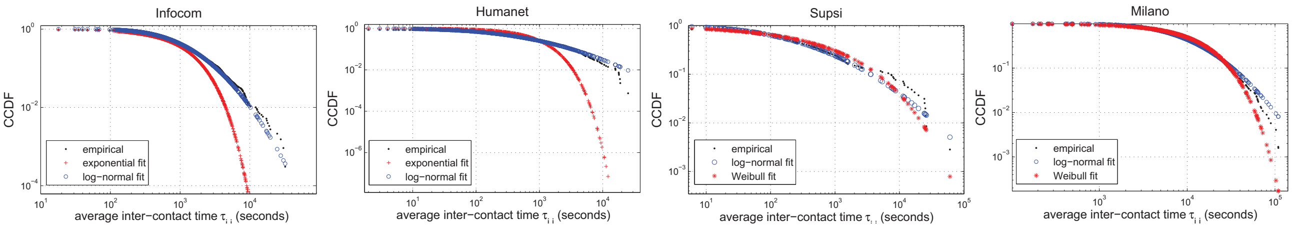
In opportunistic networking, accurately characterising inter-contact times between nodes is crucial for evaluating system performance. In earlier works, the common approach has been to look at aggregate distributions. However, recent studies such as [2], [10] indicate the risk of treating aggregate distributions as representative of pair-wise distributions. The authors in [10] prove that aggregating various pair-wise distributions can lead to false conclusions on the characteristics of node interactions on a pair level. We confirm this in the following section.

## 3.3 Experimental evaluation

In this section, we assess the capability of the homogeneous epidemic model to capture the process of content spreading in real-life scenarios.

We simulated four scenarios by replaying the pre-processed traces in 3.1. For each of the traces, we choose a single day when the nodes were most active, seen as the number of contacts recorded during that day. The reason for this is twofold: first, to observe the spreading we needed enough interaction between users, and the second, some nodes were missing from the traces during multiple days. *Humanet* trace is only one day long; for *Infocom* and *Milano* we choose the first day and for *Supsi* the eleventh. The number of nodes during those days was 72, 52, 19 and 32, for *Infocom*, *Humanet*, *Supsi* and *Milano*, respectively. The spreading works as follows. We start by infecting a single node and evaluate the time until it infects all other nodes. The same process is repeated for all the nodes in the trace. To account for the daily variations, nodes spread new content every hour during the active part of the day. We consider only the simulation runs when all the nodes were eventually infected and find the average overall delivery times and individual delivery times. Cumulative distribution functions (CDFs) for individual delivery times are plotted in Fig. 2(a).

The analytic model is an efficient tool to estimate the system performance; its simplicity stems from the fact that it requires only two input parameters: the number of nodes in the network and the node inter-contact rate. In order to validate the analytic model, we compute the same metrics,



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

**Figure 3: CCDF of average pair-wise inter-contact times: (a) Infocom, (b) Humanet, (c) Supsi, (d) Milano.**

overall and individual delivery times given by formulas (1) and (2), and assume that node interactions can be described by the aggregate inter-contact time distributions. The inter-contact rates are reciprocal to average inter-contact times. Fig. 2 (b, c) depicts the simulation results and the computed delivery times (denoted by model$^a$).

We observe large discrepancies between simulation results and the delivery times predicted by the model. The average overall delivery times obtained from the simulations are: 359, 74, 275 and 278 minutes; while the model predicts delivery times: 316, 54, 118 and 2012 seconds. Clearly, the model does not match any of the scenarios, and it underestimates the overall delivery time for the *Supsi* trace by three orders of magnitude. With respect to the average individual delivery times, the situation is similar; the simulation yields: 2089, 993, 4264 and 4291 seconds, and the model gives: 160, 28, 62 and 1039 seconds.

The explanation for this lies in several factors:

1. The aggregate inter-contact time distribution is not representative of the pair-wise inter-contact distributions in any of the examined traces. This can be seen from the aggregate distributions for *Humanet* and *Supsi* in Fig. 1. Their average inter-contact times are relatively short (order of minutes), while the distributions are long tailed. This is due to finite duration of the traces: pairs that meet more frequently will contribute more samples of their inter-contact times.
2. Many node pairs never meet. In some scenarios, we see that even the most "social" nodes meet very few other nodes and hence, all of their inter-contact times are not observable.
3. To estimate spreading times, we calculated the delivery times by averaging only over those simulations in which all the nodes were infected. In theory, the average overall delivery times would be infinite, since some nodes never get infected.

We investigate the first two findings in further detail in the following sections.

### 3.4 Pair-wise inter-contact time distributions

In all four traces, we have seen that the aggregate inter-contact time distribution does not give a complete view of the contact patterns in the network. Thus, we look at inter-contact times on a node-pair level. However, fitting different distributions for each node pair would lead to an intractable model. Hence, on a node-pair level some approximation is usually assumed. The important is that all distributions have exponential decay or at least exhibit exponential tails. For each pair of nodes in a trace, we find the average inter-contact time and plot the distributions of these average times in Fig. 3. For all the traces, log-normal distribution seems to be a good fit; we observe that tails of empirical data are bounded by log-normal and Weibull curves. We then applied curve fitting with log-normal and estimated the average inter-contact times. The average values are given in Tab. 1, denoted by $\bar{\tau}_{i,j}$. By plugging these values in the formulas, we find that the delivery times are still underestimated, although they yield better estimates than in the case of inter-contact rates calculated from the aggregate distributions. The overall delivery times are: 591, 256, 882 and 3736 seconds, and the average individual delivery times are: 300, 130, 465 and 1928 seconds (model$^b$ in Fig. 2 (b, c)).

**Table 1: Estimated average inter-contact times.**

| | Infocom | Humanet | Supsi | Milano |
|---|---|---|---|---|
| $\bar{\tau}_{ag}$ | 2347 | 312 | 323 | 7996 |
| $\bar{\tau}_{i,j}$ | 4386 | 1473 | 2397 | 14841 |
| $\bar{\tau}^C_{i,j}$ | 15187 | 16346 | 4495 | 66445 |

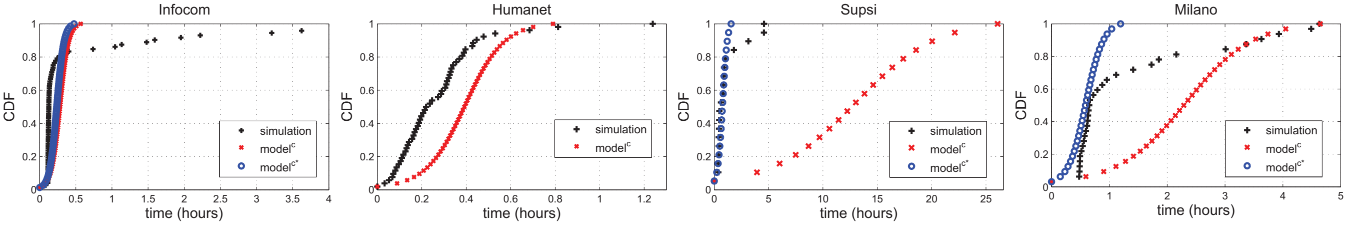
*all times are given in seconds

Clearly, node contacts are too heterogeneous: notice from Fig. 3 that average inter-contact times for different pairs differ by two orders of magnitude. Still, we want to examine if, by simply treating the traces in a different way, we can improve the estimation using the homogeneous model.

### 3.5 Compensating for the missing contacts

In a network of $N$ nodes, there are $\binom{N}{2}$ node pairs, and each pair can generate different pair-wise inter-contact time distributions. Our idea is to model contact patterns with an average inter-contact time for each node pair. Then, contact patterns in the network can be described with the *contact matrix* $\mathbf{T}$=$[\bar{\tau}_{i,j}]$, where $\bar{\tau}_{i,j}$ is the average inter-contact time for a pair of nodes $(i, j)$. $\mathbf{T}$ is a symmetric, zero-diagonal matrix since $\bar{\tau}_{i,j} = \bar{\tau}_{j,i}, \forall(i, j)$. Thus, to describe a network we need $n(n-1)/2$ matrix elements, but many node pairs are missing from the traces. For example, only 30% of all node pairs in the *Supsi* trace are observable. This raises the question how to model interaction between those node pairs and to fill in the missing elements of the contact matrix.

We use the following method: assume that all the nodes $i$ and $j$, whose contact is not captured in the processed trace, meet with some average inter-contact time $T_m$. First, we set the value $T_m$ to be equal to the duration of the entire trace: 36, 9, 103 and 62 hours. Then, we calculated the average inter-contact times for the traces by averaging over the elements of the contact matrix; these values are given in Tab. 1, denoted by $\bar{\tau}^C_{i,j}$. The results for delivery times are plotted in Figs. 2 (b, c) and 4 (model$^c$). Fig. 2 shows that the proposed method is unable to accurately estimate the delivery times in any of the scenarios. In case of the *Infocom* trace, the analytic model underestimates both overall and the individual delivery times, while for the *Supsi* trace, both times are overestimated. For the *Milano* trace, the model gives good estimation of the overall delivery time, while it



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

**Figure 4: CDF of infected population: (a) Infocom, (b) Humanet, (c) Supsi, and (d) Milano trace.**

significantly overestimates the individual delivery time. The only scenario were we observe a fairly good approximation for both delivery times is the *Humanet* scenario. However, the inconsistency of the method makes it unsuitable for use on an arbitrary trace, when the properties of the trace are not known a priori. This implication is also evident from Fig. 4; the model does not capture the evolution of the epidemic process. We also tested if filling the contact matrix with a different value for $T_m$ would give a better fit. Curves in Fig. 4 (a, b, d), denoted by model$^{c^*}$, correspond to the cases where inter-contact times $T_m$ are equal to the length of a work day (around 9 hours). Although the curves in Fig. 4 (c, d) show asymptotic behaviour in the beginning, it cannot be estimated when the simulated spreading starts to deviate from the model. For example, an accurate estimation of the time until a certain fraction of the network is infected, e.g. 80% of nodes in the *Milano* scenario, would not be possible by using this model.

The usefulness of the homogeneous model is in its simplicity, as it requires only two input parameters. However, we conclude that the homogeneous model is not accurate enough to be used for studying epidemic spreading in general, and we show methodologically, what recent studies use as a starting point and assumption but without proving, that node heterogeneity cannot be neglected when evaluating the network performance.

# 4. RELATED WORK

This paper mainly relates to the performance evaluation of epidemic content spreading. Epidemic models, adopted from the field of mathematical biology, are widely used in networking to study spreading of messages. Our study focuses on a stochastic Markov model for epidemic content distribution in opportunistic networks, proposed in [8]. A Markov model was also used in [6] to model the message delay in ad hoc networks until a specific destination was reached. The other line of work in stochastic modeling uses transient analysis of random graphs, as in [14], where the hop-limited broadcasting of messages was analysed. Studies such as [7], [15] use ordinary differential equation models, and consider the epidemic spreading process as a fluid flow. Common for all these works is that they assume homogeneous system. Heterogeneity is introduced by separating network nodes into multiple mobility classes in [9], [13], or modeling completely heterogeneous networks, as in [11]. However, it is debatable whether using these analytic models gives enough insight over simulations to account for their complexity.

# 5. CONCLUSION

We empirically evaluated the content delivery times by using four mobility datasets, chosen to represent a small system of pedestrians, moving in a relatively bounded area, and compared the empirical results with analytic model. We proposed three methods of treating the statistical data obtained from the traces. Our main finding is that a homogeneous model is unable to accurately capture the epidemic process in real-life scenarios and our future work will aim at modeling epidemic spreading in heterogeneous systems by using stochastic models.

# 6. REFERENCES

[1] J. M. Cabero, V. Molina, I. Urteaga, F. Liberal, and J. L. Martin. CRAWDAD data set Tecnalia Humanet (v. 2012-06-12), June 2012.

[2] V. Conan, J. Leguay, and T. Friedman. Characterizing pairwise inter-contact patterns in delay tolerant networks. In *Proc. Autonomics '07*, pages 19:1–19:9, Brussels, Belgium, 2007.

[3] D. J. Daley and J. Gani. *Epidemic modelling: an introduction*. Cambridge United Kingdom: Cambridge University Press, 1999.

[4] A. Förster, K. Garg, H. A. Nguyen, and S. Giordano. On context awareness and social distance in human mobility traces. In *Proc. ACM*, MobiOpp '12, New York, NY, USA, 2012.

[5] S. Gaito, E. Pagani, and G. Rossi. Fine-grained tracking of human mobility in dense scenarios. In *Proc. SECON*, 2009.

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

[7] Z. J. Haas and T. Small. A new networking model for biological applications of ad hoc sensor networks. *IEEE/ACM Trans. Networking*, 14(1):27–40, Feb. 2006.

[8] O. R. Helgason, F. Legendre, V. Lenders, M. May, and G. Karlsson. Performance of opportunistic content distribution under different levels of cooperation. In *European Wireless Conference (EW)*, pages 903–910, 2010.

[9] C.-H. Lee and D. Y. Eunt. Heterogeneity in contact dynamics: helpful or harmful to forwarding algorithms in DTNs? In *WiOPT'09*, Piscataway, NJ, USA, 2009.

[10] A. Passerella and M. Conti. Characterising aggregate inter-contact times in heterogenous opportunistic networks. In *IFIP Networking*, pages 301–313, 2011.

[11] A. Picu, T. Spyropoulos, and T. Hossmann. An analysis of the information spreading delay in heterogeneous mobility DTNs. In *Proc. IEEE WoWMoM*, pages 1–10, 2012.

[12] J. Scott, R. Gass, J. Crowcroft, P. Hui, C. Diot, and A. Chaintreau. CRAWDAD trace cambridge/haggle/imote/infocom2006 (v. 2009-05-29), May 2009.

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

[14] M. Vojnovic and A. Proutiere. Hop limited flooding over dynamic networks. In *Proc. IEEE INFOCOM*, 2011.

[15] X. Zhang, G. Neglia, J. Kurose, and D. Towsley. Performance modeling of epidemic routing. *Elsevier Comput. Networks*, 51(10):2867–2891, 2007.