General and mixed linear regressions to estimate inter-contact times and contact duration in Opportunistic Networks

Carlos Borrego[a,*], Enrique Hernández-Orallo[b], Naercio Magaia[c]

[a] *Department of Information and Communications Engineering*
Autonomous University of Barcelona, Barcelona, Spain
[b] *Departamento de Informática de Sistemas y Computadores.*
Universitat Politècnica de València, València, Spain
[c] *LASIGE, Faculdade de Ciências*
Universidade de Lisboa, Lisbon, Portugal

Abstract

In the context of Opportunistic Networking (OppNet), routing and delivery algorithms used for content dissemination employ different metrics to perform accurate decisions. It has been shown that of these metrics, the inter-contact time and the contact duration are very useful for characterising OppNet scenarios. In this article, we show that the exponential moving averages of the historical values of these metrics are correlated with future observed values, in addition to also being good estimators for them. Moreover, we go a step further to investigate how to locally, from the OppNet node perspective, improve the estimations for these metrics by defining two novel estimation functions. These estimation functions are based on two different linear models: a general regression model and a mixed regression model, where future values of the studied metrics are explained in terms of their corresponding exponential moving averages. Experimentation using real mobility traces from well-known OppNet scenarios show that our estimation functions greatly reduce the estimation error of the future values of both metrics when compared to representative state of the art proposals.

Keywords: Opportunistic Networks, inter-contact time, contact duration, regression models, mixed models

1. Introduction

Opportunistic Networking (OppNet) [1] is an open research field of computer networks that studies networks where mobile nodes communicate with each other

*Corresponding author
Email addresses: carlos.borrego@uab.cat (Carlos Borrego), ehernandez@disca.upv.es (Enrique Hernández-Orallo), ndmagaia@ciencias.ulisboa.pt (Naercio Magaia)

even when there is no end-to-end connectivity between them. As mobile devices get smarter, OppNet has emerged as a solid network solution that allows different applications to be deployed when under these network conditions.

One of the most challenging issues in OppNet is the routing and delivery decisions. Routing is the node decision as to whether or not to forward a certain message upon contacting another node. The delivery decision determines whether a node should be considered to be the destination of the message. Unlike in connected networks, OppNet routing and delivery decisions are not trivial because of OppNet's dynamical properties when it comes to a node's movement models.

In the OppNet literature, many different network metrics have been used to help with these routing and delivery decisions. Among these metrics, inter-contact time and node contact duration have shown, in a wide range of publications, to be extremely useful as they are metrics that effectively characterise the network's node behaviour [2]. While the inter-contact time metric measures the time between two successive contacts, the contact duration metric determines how long two nodes are in contact.

Examples of the benefits of these metrics in the literature are numerous. For instance, many already published proposals have shown that the contact duration distribution measured by each node is a very efficient metric for helping to increase the probability of simultaneous forwarding by deliberately postponing message forwarding [3]. In the very same way, the inter-contact time has proven to be a very efficient metric for predicting the number of future contacted nodes in a certain period of time. These kinds of predictions are very useful for complex message delivery decisions, such as in proposals like [4].

In order to estimate future values of the inter-contact time and contact duration metrics, the usual approach is to have these nodes fit these metrics to known distributions, such as exponential or power-law distributions, as in proposals like [5]. Another way of estimating these metrics is to allow OppNet nodes to build and locally store previous metric average values from the past and use these averages as estimators for future metric values. For this purpose, statistical exponential moving averages ($ewma$) are very useful because with a single average value both old values and recent values are considered. This is obtained by applying weighting factors which decrease exponentially and never reach zero.

In this article, we confirm that these exponential moving averages of historical values of the two studied metrics (inter-contact time and contact duration) are tightly correlated with future observed values, while also being good estimators for them. In addition, we take a further step forward and focus on investigating how to locally, from the OppNet node perspective, improve the estimations for these metrics by defining two novel estimation functions that use their corresponding exponential moving averages. These estimation functions are created by defining two different linear models: a general linear regression model and

a linear mixed regression model, where the explained variables are the future values of the studied metrics and the explanatory variables their corresponding exponential moving averages. In order to achieve this, we propose that the samples needed to define the proposed linear models be collected by the very same OppNet nodes, and, in the case of the mixed models, even share them upon encountering other nodes.

Specifically, the contributions of this article can be summarised as follows:

- A first estimation function based on a general linear regression model to predict the inter-contact time and the contact duration for OppNet.
- A second estimation function based on a mixed linear regression model to predict both metrics, based on node cooperation.
- We use real mobility traces to validate the correctness and performance of the two proposed estimation functions.

The paper starts with Section 2, which contains some background information on exponential averages and linear models. Then, in Section 3, we examine the state of the art of OppNet proposals that analyse the inter-contact time and contact duration metrics. Next, we provide a full description of our estimation models in Section 4. The paper continues with Section 5, where we present a comparison of our proposal with others from the state of the art. Finally, Section 6 contains the conclusions we have drawn from this work.

2. Background

In this section, we present some background information needed to understand our proposal. This will be a brief introduction on moving averages, general linear regressions, and linear mixed models.

2.1. Moving averages

An exponential moving average (ewma) is a very useful statistic tool for averaging. It is a lightweight (in terms of storage requirements) moving average that, in an adaptive way, takes into account both historical information from the past and new recently measured information. A general ewma is calculated every time a new measure is obtained in the following way:

$$ewma_{new} = ewma_{old} + \sigma \times (lastmeasured - ewma_{old}) \quad (1)$$

where $ewma_{new}$ is the new average value, $ewma_{old}$, the old value, $lastmeasured$ is the last value measured and $\sigma \in [0..1]$ a constant to give different weights to historical information and new information.

This way of averaging information using ewma has been widely employed in OppNet because of these characteristics. For example, in [6], the authors propose an OppNet routing protocol that uses an ewma-based scheme in order to keep nodal contact probabilities updated. Additionally, in [7, 8], the authors use ewma to estimate future times between forwarding actions.

2.2. General linear regressions and mixed linear models

Linear regression [9] is a basic statistical tool for predictive purposes. It is extremely useful for finding a relationship between two or more continuous variables. It studies whether a set of variables are good for predicting a certain outcome, also called the dependent variable. Given a random sample collection $(X_i, Y_i), i = 1, \ldots, n$, the simplest form of a regression model with one dependent and one independent variable is defined by the equation of the general linear regression:

$$Y_i = \alpha + \beta \times X_i + \epsilon_i, \epsilon_i \sim N(0, \sigma^2) \tag{2}$$

where Y_i is the explained response (or dependent) variable, X_i the explanatory (or independent) variable, α is called the intercept value (the value of the dependent value when the explained equals 0) and β the slope value. The information that cannot be explained by this model are the residuals ϵ_i, which are assumed to be normally distributed with expectation 0 and variance σ^2.

The model coefficients can be obtained as [10]:

$$\alpha = \frac{(\sum Y_i) \times (\sum X_i^2) - (\sum X_i) \times (\sum X_i \times Y_i)}{n \times (\sum X_i^2) - (\sum X_i)^2} \tag{3}$$

Additionally, β is obtained in the following way:

$$\beta = \frac{n \times (\sum X_i \times Y_i) - (\sum X_i) \times (\sum Y_i)}{n (\sum X_i^2) - (\sum (X_i))^2} \tag{4}$$

Linear mixed models [11] are an extension of simple linear models that also take into account both the variation explained by the independent or explanatory variables of interest, also called fixed effects, as well as the variation not explained by the independent or response variables of interest, also called random effects. They are especially useful when there is non-independence in the data, such as when the data is hierarchically structured.

The mixed linear model where information is structured into different groups $(i = 1, \ldots, N)$ and every group i contains n_i observations can be defined as:

$$Y_i = X_i \times \beta + Z_i \times b_i + \epsilon_i \tag{5}$$

where, again as in the linear regression model, X_i contains the explanatory variable, but Y_i contains the explained variable for the group i, and where N is the number of groups. There are two components in this model: the fixed term

$(X_i \times \beta)$ and the random one $(Z_i \times b_i)$. The component $Z_i \times b_i$ contains the effect on the model for every group. This means that every group is allowed to have a different $Y : X$ relationship. However, the β factor is applied to all of the groups. In order to represent X_i and Z_i, two matrices of dimension $n_i \times p$ and $n_i \times q$, are defined respectively, where n_i is the number of observations in Y_i (the number of observations per group), p the number of explanatory variables in X_i and q the number of explanatory variables in Z_i.

A special case of mixed linear models that will be used in this proposal is the random intercept model, where the model intercept may change per group. The model is defined as:

$$Y_{ij} = \alpha + \beta_1 \times Z_i + \beta_2 \times X_{ij} + \epsilon_{ij} \tag{6}$$

where Y_{ij} is the explained variable, α is the independent term, β_2 is the fixed effect associated to the fixed X_{ij} variable, and β_1 is the fixed effect associated to the random Z_i variable, which has as many levels as there are different sample groups. The random variable is included in the model assuming that the variations around the intercept for each member of Z_i is normally distributed with a certain variable. This mixed effect is modelled as:

$$Y_i = X_i \times \beta + Z_i \times b_i + \epsilon_i \tag{7}$$

In a matrix notation, the same model can be defined as:

$$\begin{bmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{bmatrix} = \begin{bmatrix} 1 & X_{i1} \\ 1 & X_{i2} \\ & \\ 1 & \vdots \\ 1 & X_{in_i} \end{bmatrix} \times \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \times \beta_i + \begin{bmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \vdots \\ \epsilon_{in_i} \end{bmatrix} \tag{8}$$

From this model, we can extract its coefficients, which are the common slope (*slope*) and a list of intercept values for the different groups ($intercept_i, i = 1, \ldots, N$). These values are the combination of the fixed effects and the variance components of the random effects, as explained in [12].

3. Related Work

In this section, we study the state of the art of Opportunistic Networking, paying special attention to two of the metrics that will be estimated in this study: inter-contact time and contact duration.

3.1. OppNet Routing and Delivery

Routing and delivery in OppNet are the decisions that set the path of a message from a source to a destination. There are several surveys that cover the different challenges routing protocols must face in OppNet. Some examples of

these surveys are [13, 14, 15, 16]. Among the different metrics used for routing and delivery decisions, for this study we focus on two of them: inter-contact time and contact duration.

3.2. Inter-contact time and contact duration aware routing/delivery algorithms

The inter-contact time between nodes in an OppNet has been proposed as a key metric for defining its performance since it has been proven to have a broad impact on message delivery reliability [2]. For this reason, the inter-contact time has been widely used as a metric for performing routing or delivery decisions. For instance, there is a wide range of social-aware OppNet routing proposals that identify important nodes in a network by using different centrality metrics. A large part of these metrics are computed taking into account the inter-contact time of the nodes: the shorter the inter-contact time is in a node, the more frequent that node will be encountered, thus increasing its popularity. For example, in [17], the authors propose a social popularity-based routing protocol named SPBR that takes into account the inter-contact time and multi-hop neighbour information.

In the context of OppNet delivery protocols, in [4], the authors propose a general delivery scheme for manycast group communications based on a mobile code. They present an application of this scheme for solving, by way of an analytical delivery method, the problem of sending a message to k and only k nodes of a heterogeneous and opportunistic network scenario that best fit a given criterion. For these purposes, the authors use for their delivery protocol the number of potential nodes belonging to a certain profile to which a message can be forwarded for a given period of time. This number is calculated by understanding the inter-contact time of every node in the network.

Analysing the duration of node contacts in OppNet, also known as contact duration, has helped propose very efficient OppNet routing, delivery, data replication schemes and caching protocols. For example, in the context of OppNet cooperative caching, Zhuo et al. [18] propose a caching protocol that takes into account the impact of the contact duration limitation on cooperative caching by deriving an adaptive caching bound at each OppNet node and analysing its specific contact patterns with other nodes. In [19], the authors propose a contact duration-aware replication protocol that operates in a fully distributed manner. In this proposal, the authors give an analytical description of the contact duration-aware data replication problem and propose a centralised solution for improving the utilisation of the storage buffers as well as the contact opportunities. In [20], the authors propose an OppNet routing protocol that avoids forwarding failure by dividing OppNet messages into smaller fragments. The authors present a mathematical model that considers the contact durations for deriving the optimal fragment size that minimises message delivery delay.

Proposals like [2, 17, 4, 18, 19] described in this section show that the inter-contact time and contact duration metrics are useful tools for making network decisions, such as the delivery or the routing decision. In turn, these network decisions aim to improve the performance of the network in terms of metrics, such as the delivery latency or the delivery ratio. In the following section, we analyse proposals that make estimations on these metrics.

3.3. Estimating inter-contact time and contact duration

As detailed above, most routing algorithms require a good estimation of the inter-contact time and contact duration to provide reliable routing decisions. Thus, the OppNet research community has proposed different ways of estimating the inter-contact time of nodes in order to make good routing or delivery decisions.

In general, users tend to follow mobility patterns, that is, humans have strong habits they follow every day, allowing future behaviour to be predicted [21]. Therefore, the main idea is to predict future encounters among nodes using the past history of encounters [22]. Concretely, this paper studies the predictability of a node's future contacts by analysing their previous contacts. The authors propose modelling the contact time series as a Poisson distribution (with the number of contacts being the Poisson events in a fixed one-hour interval), showing that the number of contacts per time unit in the future can be efficiently predicted. This model has been evaluated and validated using a trace recorded in the Politehnica University of Bucharest, in addition to other known traces. Based on this contact prediction model, along with other social criteria, the authors propose in [23] a new opportunistic routing algorithm (SPRINT), which improves the delivery ratio of messages as compared to traditional social-based routing approaches.

A further improvement is JDER [24], a new probabilistic forwarding scheme also based on the history of previously encountered nodes, one that determines the network cut-nodes, which will increase the probability of reaching the destination nodes. Another approach is to use a multi-objective model and decision tree based mechanism for optimising data dissemination under different target performances [25].

Additionally, other works like [26] estimate OppNet the inter-contact time of nodes by analysing real mobility traces. Studies like [5] have proven that the aggregate inter-contact times distribution in OppNet can be fitted to an exponential distribution for scenarios where nodes are vehicles. Instead, in [27], a similar exponential model for scenarios where nodes are human is proposed. Some other works propose compound distributions to explain the inter-contact time. For example, in [26], the authors prove that there is a characteristic time, approximately half a day, beyond which the distribution of the inter-contact

time follows a power law distribution. After this time, the distribution decays exponentially.

Nevertheless, this established approach of using the inter-contact times between pairs has been shown to achieve non-representative characterisations [16]. Using the individual nodes' inter-contacts instead can lead to more precise estimations. Furthermore, the aggregate inter-contact times distribution can only be representative of the individual distributions when all nodes contacts patterns are supposed to be the same (the *homogeneous network* assumption) [15]. Therefore, considering that opportunistic networks are heterogeneous, the individual nodes' inter-contact time will lead to better estimations as proposed in this paper.

Summing up, in this section, we have presented the state of the art of inter-contact time and contact duration estimation on OppNet. To the best of our knowledge, there are very few proposals using linear models in OppNet. In [28], the authors show from an analytical perspective that inter-contact times in OppNet can be approximated as exponentially distributed in certain mobility models. In order to prove this, they use linear regression analysis. Additionally, in [29], by using real mobility traces, the authors of this study classify users as being vagabonds or socials according to their social behaviour. They conclude, using linear regressions as statistical tools, that the effectiveness of OppNet message dissemination predominantly depends on vagabonds because they outnumber socials. However, as far as we know, there is no proposal in OppNet literature that uses ewma as an explanatory variable for creating linear general or mixed models for the purpose of estimating future values of inter-contact times and contact duration. In the following section, we explain our estimation proposal.

4. Network Model

In this section, we present our two proposed models for estimating future values for the inter-contact time (*ict*, from this point on) and the contact duration (*cd*, from this point on) metrics. First, in Section 4.1, we describe the functional architecture of the proposed estimation system and the required data. Then, in Section 4.2, we present our first estimation function based on a simple linear regression model. Finally, in Section 4.3, we complement this estimation function by proposing a second model based on linear mixed models. We provide in Algorithm 1 a list of the procedures needed to build the two models proposed. Along with the network model description, we will reference the related lines of the algorithm for the sake of clearness.

4.1. Functional architecture and dataset

The final objective of the proposed linear estimation models is to provide a module (or object) that can estimate the *ict* and *cd* values from the collected

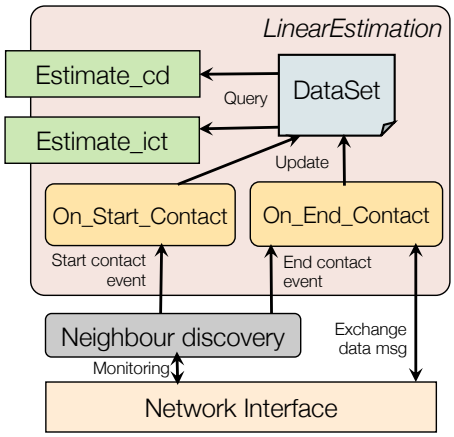


Figure 1: Architecture of the estimation module.

information of previous contacts, which can be used, for example, by the routing protocols. A possible functional structure of this module (*LinearEstimation*) is depicted in Figure 1. We also provide an algorithmic implementation of the main methods of this module in Algorithm 1.

The main interface of the *LinearEstimation* module consists of two methods for obtaining the *cd* and *ict* values based on the current information. This information is updated when a new contact starts or ends. In OppNet, two nodes contact each other if, by means of any neighbour discovery protocol, they find that they are within their communication range. Therefore, this *neighbour discovery protocol*[1] module is required to detect new contacts (and their durations) and then to notify the *LinearEstimation* module. Therefore, when a connection is detected, the *On_Start_Contact* method is called to calculate the current measured (ict_m) as (see line 3 from Algorithm 1):

$$ict_m = time.now() - lasttimecontacted \tag{9}$$

where $time.now()$ is the current time and *lasttimecontacted* is the last time the node contacted any other node[2]. In the very same way, when a contact ends, the *On_End_Contact* method is called and the current *cd* value updated (line 14 from Algorithm 1):

$$cd_m = time.now() - startcontact \tag{10}$$

where *startcontact* is the time when the current contact with the *node* started. Note that the global variable *startcontact* is updated when a new contact starts,

[1]This module is provided by every OppNet implementation and it is outside of the scope of this paper.

[2]Note that for the first contact we cannot obtain an ict value as there is no previous contact.

Algorithm 1 Functions for LinearEstimation module.

```
1: procedure On_Start_Contact(node)
2:     if not first contact then
3:         ict_m = time.now() - lasttimecontacted
4:     end if
5:     startcontact = time.now()
6:     ewma_ict = ewma_ict + σ_ict × (ict_m − ewma_ict)
7:     dataset.add_ict(ict_m,ewma_ict,wframe)
8:     if mixed then
9:         dataset_2 = exchangedata(dataset)
10:        dataset.update(dataset_2)
11:    end if
12: end procedure
13: procedure On_End_Contact(node)
14:    cd_m = time.now() - startcontact
15:    lasttimecontacted = time.now()
16:    ewma_cd = ewma_cd + σ_cd × (cd_m − ewma_cd)
17:    dataset.add_cd(cd_m,ewma_cd,wframe)
18: end procedure
19: function Estimate_ict return ict
20:    if mixed then
21:        (int,slope,ewma_ict)=mxmodel_ict(dataset,node)
22:        ict = int + slope × ewma_ict
23:    else
24:        (α,β,ewma_ict)=lnmodel_ict(dataset)
25:        ict = α + β × ewma_ict
26:    end if
27: end function
28: function Estimate_cd return cd
29:    if mixed then
30:        (int,slope,ewma_cd)=mxmodel_cd(dataset,node)
31:        cd = int + slope × ewma_cd
32:    else
33:        (α,β,ewma_cd)=lnmodel_cd(dataset)
34:        cd = α + β × ewma_cd
35:    end if
36: end function
```

and *lasttimecontacted* is updated when a contact ends (lines 5 and 15 from Algorithm 1).

Additionally, the node keeps an exponential moving average of the two studied metrics that get updated in the *On_Start_Contact* and *On_End_Contact* in the following way (lines 6 and 16 from Algorithm 1):

$$\begin{aligned} ewma_{ict} &= ewma_{ict} + \sigma_{ict} \times (ict_m - ewma_{ict}), \\ ewma_{cd} &= ewma_{cd} + \sigma_{cd} \times (cd_m - ewma_{cd}), \end{aligned} \tag{11}$$

where $\sigma_{ict} \in [0,1]$ and $\sigma_{cd} \in [0,1]$ are the exponential moving average weight constants for giving different weights to historical measured values and new ones.

In order to estimate the *ict* and *cd* values, every node *node* locally builds a dataset DS_{node}, that contains a list of sample units:

$$DS_{node} = \{ds_1, ds_2, ..., ds_{|DS_{node}|}\} \tag{12}$$

In turn, every element ds_d ($d = 1, \ldots, |DS_{node}|$) that belongs to dataset DS_{node} consists of five elements:

$$ds_d = [type_d, measured_d, ewma_d, time_d, sampler_d] \tag{13}$$

where $type_d$ is the type of metric (*ict* or *cd*), $measured_d$ is a measured value for any of the two studied metrics (calculated using equations 9 and 10), $ewma_d$ is the value for their corresponding exponential average (calculated using equation 11), $time_d$ is the time when the metric was measured and $sampler_d$ is the identification of the node that sampled the measured value[3]. The local dataset is updated with new *ict* entry when a new contact starts (line 7 from Algorithm 1) and with a new *cd* when the contacts ends (line 17 from Algorithm 1).

In order to reduce the size of these datasets, we define a temporal window frame of size *wframe*. That is, old elements are not considered from the current dataset DS_{node} if the entries are older than *wframe*:

$$\begin{aligned} &DS_{node}(wframe) = \{ds_1, ds_2, ..., ds_{|DS'_{node}|}\} \subseteq DS_{node} \\ &| \ \forall ds_d = type_d, measured_d, ewma_d, time_d, sampler_d, \\ &ds_d \in DS_{node} \ \Rightarrow \ time.now() - time_d < wframe. \end{aligned} \tag{14}$$

where $time.now()$ is the current time and *wframe* indicates a period of time during which samples may be considered. This also reduces the size of the messages exchanged. As we will see in Section 4.3, in our mixed model, when a contact occurs, both nodes exchange their datasets (using function *exchangedata*, line 9 from Algorithm 1). Therefore, when two nodes $node1$ and $node2$ meet, with datasets DS_{node1} and DS_{node2}, respectively, they update their datasets:

$$\begin{aligned} DS_{node1} &= DS_{node1} \cup DS_{node2} \\ DS_{node2} &= DS_{node2} \cup DS_{node1} \end{aligned} \tag{15}$$

This procedure is performed in both contacted nodes using *dataset.update*, as shown in the Algorithm 1, line 10.

Finally, the impact of the *wframe* variable will be analysed in Section 5, to see its performance on our proposal.

4.2. Linear regression model

In this section, we introduce the first estimation function, based on a simple regression model.

[3]Note that a sampler node may be a different node from the one that keeps the sample unit.

4.2.1. Model definition

After ending a contact with its n^{th} contacted node, any given node in the network collects a dataset of $n-1$ sample units for the pair of values $ict_i : ewma_{icti}, i = 1, \ldots, n-1$ and n samples for $cd_i : ewma_{cdi}$, $i = 1, \ldots, n$ following equations 9 - 11. Given these previous values, we propose a model for explaining the values of ict_{i+1} in terms of $ewma_{icti}$, and the values of cd_{i+1} in terms of $ewma_{cdi}$ using a bivariate linear regression model as:

$$\begin{aligned} ict_{i+1} &= \alpha_{ict} + \beta_{ict} \times ewma_{icti} + \epsilon_{icti} \\ cd_{i+1} &= \alpha_{cd} + \beta_{cd} \times ewma_{cdi} + \epsilon_{cdi} \end{aligned} \tag{16}$$

where

$$\epsilon_{icti} \sim N(0, {\sigma_{ict}}^2), \epsilon_{icti} \sim N(0, {\sigma_{cd}}^2) \tag{17}$$

Thus, ict_{i+1} and cd_{i+1} are the response variables explained by $ewma_{icti}$ and $ewma_{cdi}$, respectively. The information not explained by these models is defined by the residuals ϵ_{ict} and ϵ_{cd}, which are assumed to be normally distributed with expectation 0 and variance ${\sigma_{ict}}^2$ and ${\sigma_{cd}}^2$, respectively.

4.2.2. Estimation function

Consequently, a node *node* that at a given time has measured and calculated m sample values can calculate the estimation for the future values of *ict* and *cd* as:

$$\begin{aligned} ict_{m+1} &= \alpha_{ict} + \beta_{ict} \times ewma_{ictm} \\ cd_{m+1} &= \alpha_{cd} + \beta_{cd} \times ewma_{cdm} \end{aligned} \tag{18}$$

where the α and β values can be obtained using equations 3 and 4 detailed in section 2. The implementation of these estimations is shown in Algorithm 1, where the *Estimate_ict* and *Estimate_cd* functions calculate first the α, β and *ewma* values using the dataset through functions *lnmodel_ict* and *lnmodel_cd* for later applying equation 18 (lines 25 and 34).

4.3. Linear mixed model

In the previous subsection, we have proposed a linear regression model for predicting future values for *ict* and *cd* in terms of the already measured ones. Every node performed independent linear regression models. In this section, we extend our proposal by allowing nodes to share their measured values for both metrics in order to obtain a more elaborate linear regression mixed model.

4.3.1. Model definition

As explained in Section 2.2, mixed models are very useful in contexts where data are clustered or hierarchically organised. In our mixed model, the information gathered is clusterised, thus creating groups. These groups are the different nodes in the network. So, as random factors, we use the qualitative variable *node*,

which represents the original node that measured the information, and, as fixed effects, we use the studied metrics and their corresponding exponential moving averages. The two expressions of random intercept models below represent our proposed mixed models.

First, we model the inter-contact time (ict_{d+1}) as a linear function of $ewma_{ictd}$, where the intercept is allowed to change per each node ($node_i = 1, \ldots, N$). In this model, $node_i$ is a factor with as many levels as nodes considered in the networks, where N is the maximum number of nodes in the network. This means that nodes group their datasets in N clusters of n_i sample units for each group:

$$\begin{aligned} &ict_{i,j+1} = \alpha_{ict} + \beta_{ict1} \times node_i + \beta_{ict2} \times ewma_{ictij} + \epsilon_{ictij} \\ &i = 1, \ldots, N; j = 1, \ldots, n_i \end{aligned} \tag{19}$$

There are two components in this model that include explanatory variables: the fixed one, $\beta_{ict2} \times ewma_{ictij}$, and the random one, $\beta_{ict1} \times node_i$, which represents the $ict : ewma_{ict}$ effect for every node. Each node is allowed to have a different $ict : ewma_{ict}$ relationship. Our model considers the type of mixed model where the different nodes have the same slope, with different intercepts per *node*. We are using *node* as a random effect, assuming that the variations on the intercept for each *node* is normally distributed with a certain variance. This means that the differences per node regarding the intercept are small. Thus, ϵ_{ictij} represents the errors that are normally distributed with covariance matrix Σ, $\Sigma = \sigma^2 \times I$.

Secondly, in the very same way, we proposed the following mixed model to estimate the contact duration:

$$\begin{aligned} &cd_{i,j+1} = \alpha_{cd} + \beta_{cd1} \times node_i + \beta_{cd2} \times ewma_{cdij} + \epsilon_{ictij} \\ &i = i, \ldots, N; j = 1, \ldots, n_i \end{aligned} \tag{20}$$

Once both mixed models are defined, as explained in Section 2.2, we can extract its coefficients, that is, the common slopes ($slope_{ict}$ and $slope_{cd}$, for the *ict* and *cd* mixed model, respectively) and a list of intercept values for the different nodes ($intercept_{ict_i}$ and $intercept_{cd_i}$, $i = 1, \ldots, N$, for the *ict* and *cd* mixed model, respectively).

4.3.2. Estimation function

Consequently, a node (*node*) that for a given time has obtained m_{ict} and m_{cd} sample values in its dataset DS_{node} may calculate the estimation for the future values of *ict* and *cd* as:

$$\begin{aligned} &ict_{m_{ict}+1} = intercept_{ict_{node}} + slope_{ict} \times ewma_{ictm_{ict}} \\ &cd_{m_{cd}+1} = intercept_{cd_{node}} + slope_{cd} \times ewma_{cdm_{cd}} \end{aligned} \tag{21}$$

where $intercept_{ict_{node}}$ and $intercept_{cd_{node}}$ are the intercept values for node *node* extracted from the *ict* and *cd* mixed models, $slope_{ict}$ and $slope_{cd}$ their common

slopes, and $ewma_{ict_{m_{ict}}}$ and $ewma_{cd_{m_{cd}}}$ the last $ewma_{ict}$ and $ewma_{cd}$ values calculated. The implementation of these estimations is shown in Algorithm 1 (lines 22 and 31), where the *Estimate_ict* and *Estimate_cd* functions calculate first the intercept (*int*), slope and *ewma* values using the dataset through functions *mxmodel_ict* and *mxmodel_cd*, to be applied later in equation 18.

5. Evaluation

In this section, we present the experiments carried out to evaluate the performance of our proposal. First, we present our evaluation methodology used in this experimentation. Then, we present the results.

5.1. Evaluation methodology

For the experiments conducted for this article, we chose five different scenarios to analyse the performance of our proposal. The physical encounters for the five different scenarios were obtained from real mobility traces from the Crawdad database[4], a community resource for collecting wireless data at Dartmouth College, United States. The five scenarios are the following:

- The first scenario, the *Info5* scenario, is based on real mobility traces, as published in [30]. These traces were retrieved during the 2005 edition of the Infocom conference over the course of 2.97 days. Contacts from these mobility traces represent 41 students carrying *iMote* platforms[5]. The number of contacts provided in these traces is 22,459.

- The second scenario, the *Cambridge* scenario, as published in [30], is based on real Bluetooth traces from students from the System Research Group of the University of Cambridge, UK, carrying small devices for six days. Additionally, some stationary nodes were placed at various points of interest. The number of contacts provided in these traces is 10,873.

- The third scenario, the *MIT* scenario, as published in [31], represents the activity information from 100 subjects at the Massachusetts Institute of Technology over the course of the 2004-2005 academic year. The number of contacts provided in these traces is 102,593.

- The fourth scenario, the *Asturias* scenario, as published in [32], contains connectivity traces extracted from GPS traces obtained from the regional

[4]Mobility traces can be found at *http://crawdad.org/*.

[5]An iMote is a simple device made by Intel Research based on a Zeevo TC2001P system-on-a-chip providing an ARM7 processor, Bluetooth connectivity and a 950mAh CR2 bakery.

Fire Department of Asturias, Spain. This data was generated by GPS devices embedded in different vehicles such as cars, trucks and a helicopter and a few personal radios. In total, 229 devices reported 3,098,642 contacts.

- The fifth scenario, the *Taxis* scenario, as published in [33], contains mobility traces from 320 taxi cabs in Rome over 30 days. The number of contacts provided in these traces is 224,588.

A summary with the general characteristics of the scenarios can be found in Table 1. Although the traces have different durations, in all cases we have only used the first 24 hours in order to compare results using equivalent time intervals.

The traces data are analysed to obtain information about *ict* and *cd* metrics. Some of these mobility traces contain information on the times when nodes start contact and end their encounters. Some others, such as the Taxis trace, contains GPS coordinates instead. For this type of traces, we assume a range of 10m to obtain the information on when the encounter starts and finishes (which can resemble a typical Bluetooth range). Once the five traces were analysed, considering only the first 24 hours, we obtained a dataframe with the following fields[6]:

- **Type:** represents the type of sample unit: *ict* or *cd*.
- **Sampler:** name of the node that has measured the sample.
- **Custodier:** name of the node that keeps the sample unit.
- **Value:** inter-contact time or contact duration sample unit (seconds).
- **Ewma:** exponential weight moving average of inter-contact time or contact duration before the sample unit was obtained.
- **Time:** time of the sample unit (seconds)
- **Scenario:** name of the scenario

For example, an entry of our dataset containing the following values:

```
Type Sampler Custodier Value EWMA Time Scenario
cd   x9      x25       30    35   450  Asturias
```

[6]Dataframe with all the traces and metric values can be requested at: http://deic.uab.es/~cborrego/linear.html

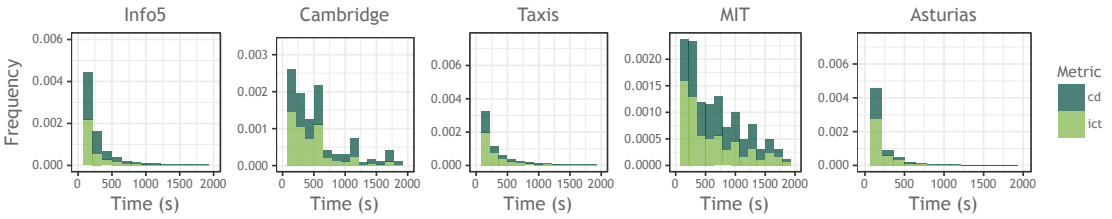


Figure 2: Inter contact time and contact duration histograms for the five different scenarios.

meaning that node $x25$ keeps a contact duration sample unit learned from $x9$ that was sampled at time 450 seconds with a value of 30 seconds and the previous calculated EWMA was 35 seconds.

The evaluation results presented in the following section have been performed by analysing this dataframe using the statistical tools R, MATLAB, and Python[7].

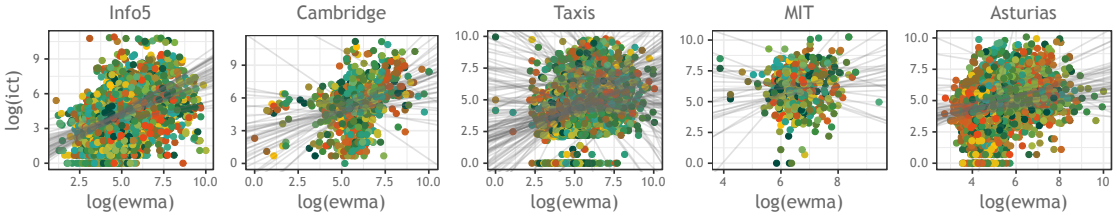
Finally, for the values of σ_{ict} and σ_{cd} in equations 11, we have used the constant value 0.7, as in classical proposals such as [34, 35, 36], and *wframe* was initially set to 24h.

5.2. Performance evaluation

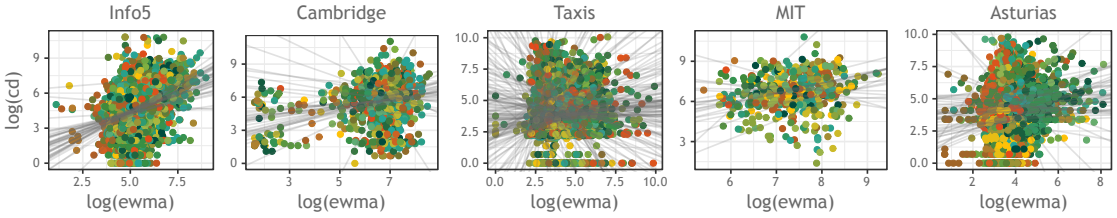
The goal of this section is to evaluate the performance of our estimation proposals by comparing them with other well-known methods for estimating *ict* and *cd*. We present the different methods that will be compared in this evaluation as follows:

- Linear model (**LM** in the figures) as explained in Section 4.2.
- Linear mixed model (**MX** in the figures) as explained in Section 4.3.
- Pareto distribution (**PD** in the figures). The estimation of the inter-contact time and contact duration is performed using a Pareto distribution, which is previously calculated using the whole contact trace, as in proposals like [37].
- Exponential distribution (**XD** in the figures). This is similar to **PD** but using an exponential distribution, as in proposals like [5]. Note that both **PD** and **XD** are *offline* methods and therefore cannot be implemented in the nodes.

[7]Source code for the statistical tools can be requested at:
http://deic.uab.es/~cborrego/linear.html



(a) Linear model $ict_{i+1} = \alpha_{ict} + \beta_{ict} \times ewma_{ict\,i} + \epsilon_{ict}$.



(b) Linear model $cd_{i+1} = \alpha_{cd} + \beta_{cd} \times ewma_{cd\,i} + \epsilon_{cd\,i}$.

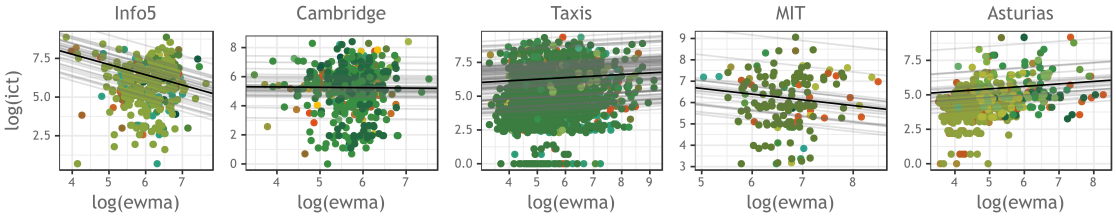
Figure 3: *ict* and *cd* general linear models after 24 hours for every node in the network. Colour circles represent metric values from the different nodes in the network. The circle colour is used to differentiate the different nodes. The grey lines depict their corresponding general linear models.

- Pareto (**PW** in the figures). The estimation of the inter-contact time and contact duration is performed using a Pareto distribution as well, but, this time is built by every node from previously collected samples.
- Exponential (**EX** in the figures). Similar to **PW** but using an exponential distribution.
- Median (**MD** in the figures). In this case, the estimation of the inter-contact time and contact duration is performed using the median from the previous samples, as in proposals like [38].
- Moving Average (**MA** in the figures). The estimation of the inter-contact time and contact duration is performed using the exponential moving average calculated from previous samples, as in proposals like [7].

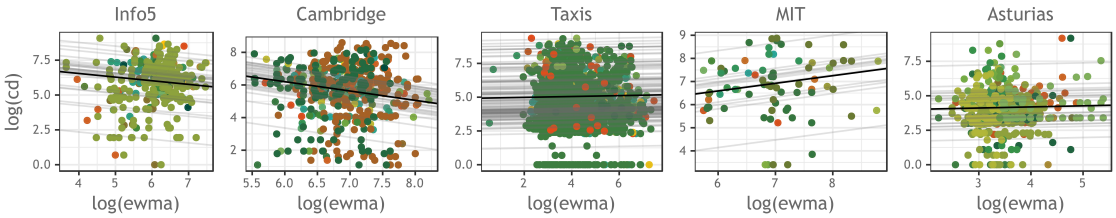
Finally, in order to evaluate the performance of the different proposals, we have defined the following *error estimation*: for a certain metric (*ict* or *cd*), we compute the average of the estimation error of a node with n_i samples of the studied metric as:

$$Error_{node\,i} = \frac{\sum_{n=1}^{n_i} |estimated_n - measured_n|}{n_i} \tag{22}$$

Before analysing the different estimation proposals, it is interesting to depict the distribution of the measured inter-contact time and contact duration for the



(a) Mixed model model $ict_{ij+1} = \alpha_{ict} + \beta_{ict_1} \times node_i + \beta_{ict_2} \times ewma_{ictij} + \epsilon_{ictij}$.



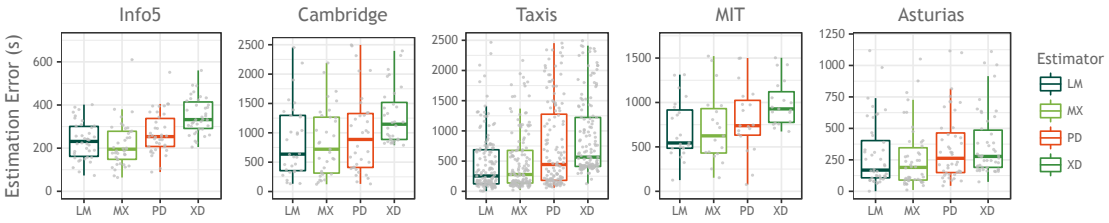
(b) Mixed model $cd_{ij+1} = \alpha_{cd} + \beta_{cd_1} \times node_i + \beta_{cd_2} \times ewma_{cdij} + \epsilon_{ictij}$.

Figure 4: *ict* and *cd* linear mixed models after 24 hours for every node in the network. Colour circles represent metric values from the different nodes in the network. The circle colour is used to differentiate the different nodes. The grey lines depict their corresponding mixed models. The black lines depict the population models.

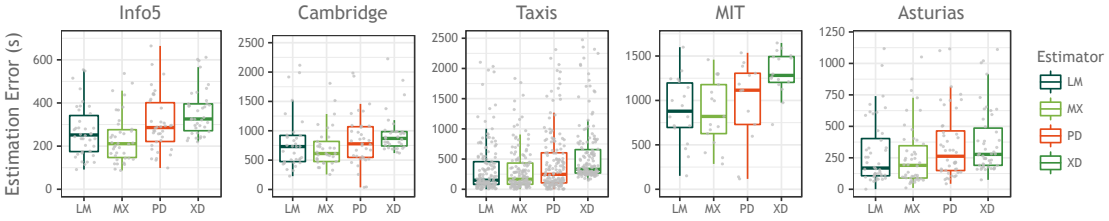
five scenarios (Figure 2). As can be seen, the histograms of the five different scenarios are very different. As can be seen by the way these metrics are distributed, their estimation is a difficult challenge.

We depict the different models obtained from the traces to show the correlation between the exponential moving averages of the historical values of the studied metrics and their corresponding future observed values. In Figure 3(a) and Figure 3(b), we depict the linear regression model introduced in Section 4.2 for the *ict* metric and the *cd* metric for every node in the five different scenarios and for 24 hours of activity. As can be seen, the nodes belonging to the Info5 scenario have similar slopes in the different models, while in the other scenarios, the slopes and intercept vary significantly for the different nodes in the network. Similarly, in Figure 4(a) and Figure 4(b), we depict the results for the mixed regression models introduced in Section 4.3. In this case, the thin lines represent the mixed models for the different nodes, while the dark line represents the population model, that is, the model that contains all the nodes.

We then proceed to compare the performance in estimating for the different methods. We divide these comparisons into two. In the first one, we compare our estimation functions against the PD and XD models, where the distribution is previously calculated. Note that these models are only considered to be a reference, since they consider the whole contact trace for estimating the values,



(a) Performance study for inter-contact time estimation.



(b) Performance study for contact duration.

Figure 5: Performance study. Our linear model (LM) and mixed model (MX) versus using classical distributions such as Pareto (PD) and Exponential (XD) for estimating the two studied metrics *ict* and *cd*.

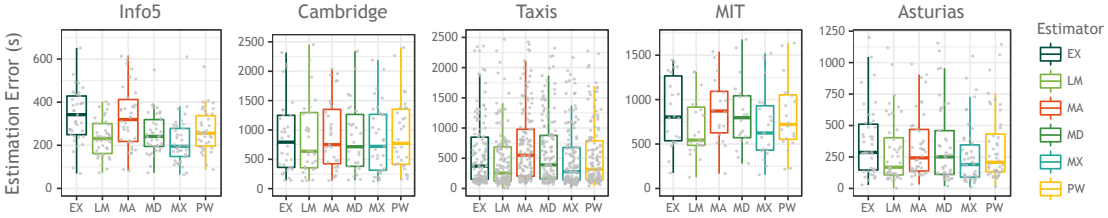
and therefore, cannot be implemented in the nodes. In Figure 5(a) and Figure 5(b), we depict the error estimation distribution, as defined in equation 22, for every node in the five different scenarios.

In the second type of comparison, we analyse the other proposals where the distributions of the behaviour of the network are ignored by the nodes in the network and decisions are taken in terms of observed events, that is, exponential moving average (MA), the historical median (MD), the exponential distribution (XP) and the Pareto distributions (PW). The results are shown in Figure 6(a) and Figure 6(b). As can be seen, our proposals –general linear model (LM) and mixed model (MX)– outperform all of the other compared proposals in terms of the estimation error for estimating both *ict* and *cd* in all of the studied scenarios.

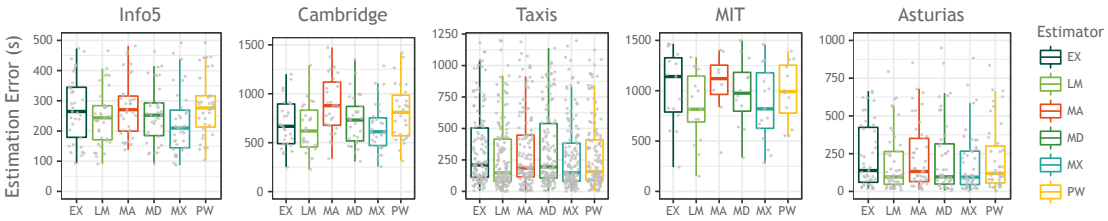
When comparing our two proposals in terms of the estimation error obtained for the *ict* metric, the scenario where the mixed model performs better than the general linear model is the Info5 scenario. Instead, for the *cd* metric, in three scenarios, our mixed model performs better in terms of estimation error than the general linear model does. These results are shown in Table 1. With data available for only 5 scenarios, it is difficult to provide a general criterion to define which proposal fits best for a general scenario.

5.3. Autocorrelation analysis and EWMA dependence

The main objective of this paper is to predict the future values of the inter-contact times (*ict*) and contact durations (*cd*) of the nodes based on the history



(a) Performance study for inter-contact time.



(b) Performance study for contact duration.

Figure 6: Performance study. Our linear model (LM) and mixed model (MX) versus built exponential distribution (EX), exponential moving average calculated from previous samples (MA), median from the previous samples (MD), built Pareto distribution (PW) for estimating the two studied metrics *ict* and *cd*.

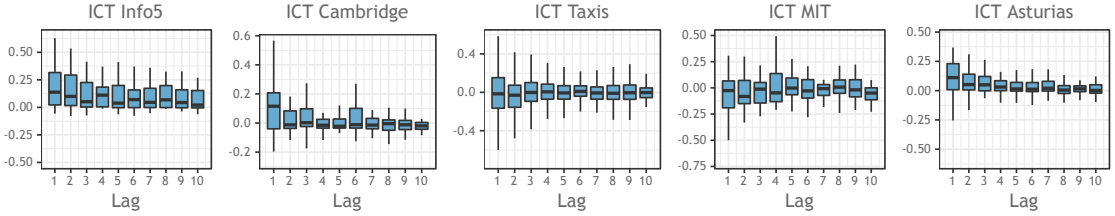
Scenario	**Nodes**	**Contacts**	**Best ict**	**Best cd**
Info5	41	10,152	MX	MX
Cambridge	51	1,216	LM	MX
Taxis	304	6,506	LM	LM
MIT	97	5,042	MX	MX
Asturias	230	12,705	LM	LM

Table 1: Main parameters of the scenarios. The number of contacts corresponds to the 24h traces extract used. Finally, the columns "Best ict" and "Best cd" represent which method from the ones proposed in this study performs better.

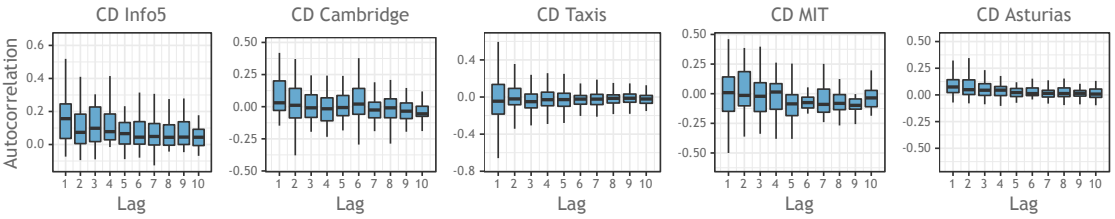
of their previous contacts. Therefore, these contacts need to have some kind of regularity. Formally, the values in the time series of nodes contacts must exhibit some degree of autocorrelation.

This autocorrelation is the basis of the EWMA estimation. Concretely, EWMA estimation is based on giving more weight to recent values of a measured times series, where the coefficient α represents the degree of weighting decrease. There is a strong dependence between EWMA and autocorrelation: highly autocorrelated time series will produce better estimations. On the contrary, time series with no autocorrelation (for example, white noise) are not predictable, and indeed, in these cases EWMA is useless.

The coefficient of correlation between two values in a time series is called the autocorrelation function (ACF). It evaluates the autocorrelation between values that are k time periods apart (known as lag k). Note that in our case, the



(a) Autocorrelation study for inter-contact time.



(b) Autocorrelation study for contact duration.

Figure 7: Autocorrelation study for the sampled values of the two studied metrics *ict* and *cd*.

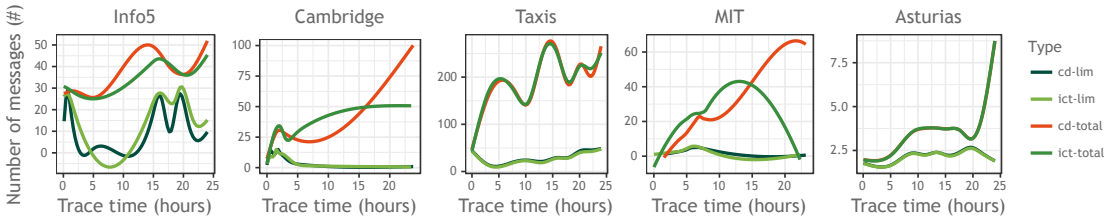


Figure 8: Number of messages exchanged as a function of time.

measurement intervals are not constant since new values are obtained when a contact occurs. Concretely, we have obtained the ACF for the *ict* and *cd* time series from the traces used in this paper. Formally, we obtained a time series for the contact of every node m of the traces: $\{ict_m\}, \{cd_m\}$.

In general, all used traces exhibit some degree of autocorrelation, as shown in the boxplots in Figure 7. These plots were obtained in the following way: for a given lag number and trace, we obtain the autocorrelation values for all the *ict* and *cd* time series of all nodes. The final result is two vectors with all the autocorrelations. We repeated this procedure for all traces and for lag values between 1 and 10. Then, the distribution of all the autocorrelations is displayed using a boxplot, showing the minimum and maximum values and the interquartile range between a first quartile (Q1=25%) and the third quartile Q3=75%). The results show that in all the traces, a high proportion of nodes have (at least) autocorrelations higher than 0.1. Furthermore, nodes exhibit higher autocorrelations for lower lag values by displaying wider interquartile ranges.

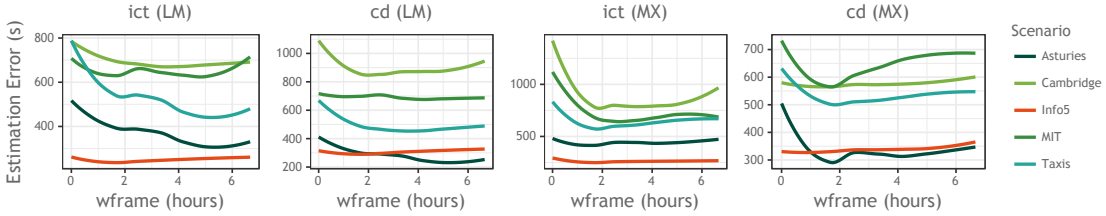


Figure 9: Error estimation as a function of the window frame for estimating the two studied metrics (*ict* and *cd*) using the general linear model (LM) and the mixed model (MX).

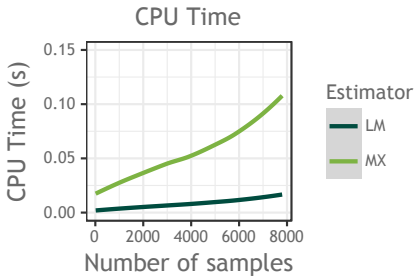


Figure 10: CPU time model calculation as a function of the number of samples.

Summing up, two main conclusions can be drawn from these results. First, there is a clear dependence between values of recent contacts, and second, that using recent values for predicting new values, as EWMA does, can obtain reasonable estimations of future *ict* and *cd* values, as is evidenced by the results obtained in the previous subsection.

5.4. Overhead evaluation and optimisation

As introduced in Section 4.3, our proposed mixed model allows nodes to share their measured values for both metrics in order to obtain a more elaborate linear regression mixed model. To learn the amount of overhead implied by the exchange of messages when implementing our mixed model, in Figure 8, we analyse the number of exchanged messages as a function of the scenario trace time. Four different types of analysis are performed for the two studied metrics: $ict-total$ and $cd-total$, when all of the information observed by nodes is shared upon node encounter and $ict-lim$ and $cd-lim$, when only messages belonging to a window frame of two hours are used (see Section 4.3 for the window frame definition). As expected, by limiting the window frame, we obtain a lower number of exchanged messages per contact. The number of exchanged messages observed seems reasonably small for OppNet. On the other hand, when not limiting the number of exchanged messages using the window frame, the number of exchanged messages could affect the performance of the network, especially in scenarios like

the Taxis scenario, where there is a large number of nodes. The value of two hours has been chosen by analysing the performance in terms of the error estimation when compared to the window frame. As can be seen in Figure 9, the *wframe* value of two hours is a good compromise value with reasonable results in terms of the error estimation for the two studied metrics, for both general and mixed models in the five scenarios.

Now, in Figure 10, we analyse the CPU overhead cost in terms of CPU seconds[8]. We see that the generic linear models are not very sensitive to the size of the samples, while mixed models are. However, the amount of time needed for calculating the model is very small, considering as well that this calculation can be computed at any time and not necessary when performing the routing or delivery action.

Summing up, although the regression mixed model has a slightly greater overhead than the linear model, mainly due to its collaborative approach, in scenarios with a lower number of nodes it can obtain the better estimation.

6. Conclusions

In this paper, we have analysed the problem of estimating future values of the inter-contact time and the contact duration in Opportunistic Networks. We have shown that exponential moving averages from historical metrics observed for both metrics are correlated with their corresponding future values.

Particularly, we have proposed two independent estimation functions based on two linear models: a general linear regression model and a linear mixed model. We have shown, by using five different real mobility traces from well-known Opportunistic Networks scenarios, that our estimation functions greatly improve the estimation of future values for the studied metrics when compared to other well-known proposals in terms of the error estimation. Furthermore, these estimation models can be easily implemented and integrated into current OppNet routing protocols, thus enabling improved performance in their routing decisions with no significant overhead.

Acknowledgements

The authors thank Teresa Rosas and Llorenç Badiella for their generous help with the statistics explanation and R programming. This work was supported by LASIGE Research Unit, ref. UID/CEC/00408/2013.

[8]Measured using a Raspberry Pi Broadcom BCM2835 SoC full HD, 700MHz Low Power ARM1176JZ-F, 512MB SDRAM, 4GB SD with Raspbian.

References

[1] L. Pelusi, A. Passarella, M. Conti, Opportunistic networking: data forwarding in disconnected mobile ad hoc networks, Communications Magazine, IEEE 44 (11) (2006) 134–141.

[2] S. Batabyal, P. Bhaumik, Improving network performance with affinity based mobility model in opportunistic network, in: Wireless Telecommunications Symposium (WTS), 2012, IEEE, 2012, pp. 1–7.

[3] K. Watabe, H. Ohsaki, Contact duration-aware epidemic broadcasting in delay/disruption-tolerant networks, IEICE Transactions on Communications 98 (12) (2015) 2389–2399.

[4] C. Borrego, G. Garcia, S. Robles, Softwarecast: a code-based delivery manycast scheme in heterogeneous and opportunistic ad hoc networks, Ad Hoc Networks.

[5] H. Zhu, L. Fu, G. Xue, Y. Zhu, M. Li, L. M. Ni, Recognizing exponential inter-contact time in vanets, in: INFOCOM, 2010 Proceedings IEEE, IEEE, 2010, pp. 1–5.

[6] H. Dang, H. Wu, Clustering and cluster-based routing protocol for delay-tolerant mobile networks, IEEE Transactions on Wireless Communications 9 (6) (2010) 1874.

[7] C. Borrego, A. Sánchez-Carmona, Z. Li, S. Robles, Explore and wait: A composite routing-delivery scheme for relative profile-casting in opportunistic networks, Computer Networks 123 (2017) 51–63.

[8] C. Borrego, J. Borrell, S. Robles, Efficient broadcast in opportunistic networks using optimal stopping theory, Ad Hoc Networks 88 (2019) 5–17.

[9] J. Neter, W. Wasserman, M. H. Kutner, Applied linear regression models, Irwin Homewood, IL, 1989.

[10] C. R. Rao, H. Toutenburg, Linear models, in: Linear models, Springer, 1995, pp. 3–18.

[11] A. Zuur, E. Ieno, N. Walker, A. Saveliev, G. Smith, Mixed effects models and extensions in ecology with R. Gail M, Krickeberg K, Samet JM, Tsiatis A, Wong W, editors, Spring.

[12] R. C. Littell, G. A. Milliken, W. W. Stroup, R. D. Wolfinger, O. Schabenberger, SAS for mixed models, SAS institute, 2007.

[13] N. Chakchouk, A survey on opportunistic routing in wireless communication networks, IEEE Communications Surveys & Tutorials 17 (4) (2015) 2214–2241.

[14] V. F. Mota, F. D. Cunha, D. F. Macedo, J. M. Nogueira, A. A. Loureiro, Protocols, mobility models and tools in opportunistic networks: A survey, Computer Communications 48 (2014) 5–19.

[15] A. Passarella, M. Conti, Analysis of individual pair and aggregate intercontact times in heterogeneous opportunistic networks., IEEE Trans. Mob. Comput. 12 (12) (2013) 2483–2495.

[16] E. Hernández-Orallo, J. C. Cano, C. T. Calafate, P. Manzoni, New approaches for characterizing inter-contact times in opportunistic networks, Ad Hoc Networks 52 (2016) 160–172.

[17] Y. Song, J. Li, C. Li, F. Wang, Social popularity based routing in delay tolerant networks., International Journal on Smart Sensing & Intelligent Systems 9 (4).

[18] X. Zhuo, Q. Li, G. Cao, Y. Dai, B. Szymanski, T. La Porta, Social-based cooperative caching in DTNs: A contact duration aware approach, in: Mobile Adhoc and Sensor Systems, IEEE 8th International Conference on, 2011, pp. 92–101.

[19] X. Zhuo, Q. Li, W. Gao, G. Cao, Y. Dai, Contact duration aware data replication in delay tolerant networks, in: Network Protocols (ICNP), 2011 19th IEEE International Conference on, IEEE, 2011, pp. 236–245.

[20] S.-H. Kim, Y. Jeong, S.-J. Han, Use of contact duration for message forwarding in intermittently connected mobile networks, Computer Networks 64 (2014) 38–54.

[21] N. Magaia, C. Borrego, P. Pereira, M. Correia, Privo: A privacy-preserving opportunistic routing protocol for delay tolerant networks, in: IFIP Networking, 2017, pp. 1–9.

[22] R.-I. Ciobanu, R.-C. Marin, C. Dobre, Interaction predictability of opportunistic networks in academic environments, Trans. Emerg. Telecommun. Technol. 25 (8) (2014) 852–864.

[23] R. I. Ciobanu, C. Dobre, V. Cristea, SPRINT: social prediction-based opportunistic routing, in: World of Wireless, Mobile and Multimedia Networks, 14th Symposium and Workshops on a, IEEE, 2013, pp. 1–7.

[24] R. Ciobanu, D. Reina, C. Dobre, S. Toral, P. Johnson, JDER: A history-based forwarding scheme for delay tolerant networks using jaccard distance and encountered ration, J. Netw. Comput. Appl. 40 (C) (2014) 279–291.

[25] D. G. Reina, R. Ciobanu, S. Toral, C. Dobre, A multi-objective optimization of data dissemination in delay tolerant networks, Expert Systems with Applications 57 (2016) 178 – 191.

[26] T. Karagiannis, J.-Y. Le Boudec, M. Vojnovic, Power law and exponential decay of intercontact times between mobile devices, IEEE Transactions on Mobile Computing 9 (10) (2010) 1377–1390.

[27] R. Groenevelt, P. Nain, G. Koole, The message delay in mobile ad hoc networks, Performance Evaluation 62 (1-4) (2005) 210–228.

[28] M. Abdulla, R. Simon, The impact of intercontact time within opportunistic networks: protocol implications and mobility models, TechRepublic White Paper.

[29] G. Zyba, G. M. Voelker, S. Ioannidis, C. Diot, Dissemination in opportunistic mobile ad-hoc networks: The power of the crowd, in: INFOCOM, 2011 Proceedings IEEE, IEEE, 2011, pp. 1179–1187.

[30] D.-G. Akestoridis, CRAWDAD dataset uoi/haggle (v. 2016-08-28): derived from cambridge/haggle (v. 2009-05-29) (Aug. 2016).

[31] N. Eagle, A. S. Pentland, CRAWDAD dataset mit/reality (v. 2005-07-01) (Jul. 2005).

[32] S. Cabrero, R. Garca, X. G. Garca, D. Melendi, CRAWDAD dataset oviedo/asturies-er (v. 2016-08-08) (Aug. 2016).

[33] L. Bracciale, M. Bonola, P. Loreti, G. Bianchi, R. Amici, A. Rabuffi, CRAWDAD dataset roma/taxi (v. 2014-07-17) (Jul. 2014).

[34] M. Mirza, K. Springborn, S. Banerjee, P. Barford, M. Blodgett, X. Zhu, On the accuracy of TCP throughput prediction for opportunistic wireless networks, in: Sensor, Mesh and Ad Hoc Communications and Networks, 2009. SECON'09. 6th Annual IEEE Communications Society Conference on, IEEE, 2009, pp. 1–9.

[35] Q. He, C. Dovrolis, M. Ammar, On the predictability of large transfer tcp throughput, in: ACM SIGCOMM Computer Communication Review, Vol. 35, ACM, 2005, pp. 145–156.

[36] M. Mirza, J. Sommers, P. Barford, X. Zhu, A machine learning approach to tcp throughput prediction, in: ACM SIGMETRICS Performance Evaluation Review, Vol. 35, ACM, 2007, pp. 97–108.

[37] A. Passarella, M. Conti, Characterising aggregate inter-contact times in heterogeneous opportunistic networks, in: International Conference on Research in Networking, Springer, 2011, pp. 301–313.

[38] 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 (6) (2007) 606–620.