Call Detail Records to Characterize Usages and Mobility Events of Phone Users

Yannick Leo[a,d,*], Carlos Sarraute[e], Anthony Busson[c,d], Eric Fleury[a,b,d]

[a] *ENS de Lyon*
[b] *Inria*
[c] *Université Claude Bernard Lyon 1*
[d] *Université de Lyon – UMR CNRS - ENS de Lyon - UCB Lyon 1 - INRIA 5668*
[e] *Grandata Labs, Buenos Aires, Argentina*

Abstract

Cellular communications are evolving quickly to constantly adapt and tolerate the load induced by the increasing number of phones. Understanding the traffic is crucial to refine models and improve experiments. In this context, one has to understand the temporal and spatial user behavior at different levels. At the user scale, the usage is not only define by the amount of calls but also by the user's mobility and type of communication. At a higher level, the BS have a key role on the flow quality. In this paper, we propose a 1-year Call Detail Records (CDR) analysis in Mexico in order to catch on usage turnovers and investigate overlooked parameters such as the call duration. Moreover, we look into handovers (switching from a station to an other one). Our study suggests that user mobility is pretty dependant to user calls.

Keywords: Mobile Traffic Analysis; Handovers; Phone User Behavior

1. Introduction

With the constant evolution of mobile technologies and digital networks, such as new generation of smartphones, and new applications, usage of cellular networks tends

☆Work partially supported by the Labex MILYON.
☆☆Work partially supported by the STIC AMSUD project UCOOL.

*Corresponding author
Email addresses: yannick.leo@ens-lyon.fr (Yannick Leo), charles@grandata.com (Carlos Sarraute), anthony.busson@ens-lyon.fr (Anthony Busson), eric.fleury@inria.fr (Eric Fleury)

to change deeply. The analysis of phone calls from real logs is thus fundamental, both from phone operators and from other stakeholders' points of view. For the operators, it gives insights on the network usage and load, and consequently on possible dimensioning issues. It also allows to adapt or propose services according to the user trends. More generally, mobile phone datasets allows to derive an analysis and statistics of human activities at a precise scale and fine level of detail. This unprecedented flow of continuous information on human activity represents a tremendous opportunity for research and real-world applications. Indeed, models or simulations that used to study and dimension cellular networks, as queuing theory for instance, need to take into account the recent evolution of networks load and may progress by considering our new observations that concern the call duration and the inter-arrivals (time between two successive calls).

In the context of a collaboration with Grandata Labs that leverages advanced research in Human Dynamics (the application of "big data" to social relationships and human behavior) to identify market trends and predict customer actions, we have access to the logs for one complete year of all calls and sms from a top-3 Mexican wireless service provider with more than seven million subscribers. It represents 90 millions of Mexicans calling each others. The availability of mobile phone datasets has opened the possibility to improve our understanding of how humans communicate, socialize, move around cities, mobilize, etc. This project plans to study these logs through different dimensions: technological, sociological and economical.

In this paper, we focus on the analysis of this trace from the network/operator point of view. We perform two analysis: one in time and one in space. First, we assess the phone usage in terms of load for different time periods, and study the distribution of the quantities that impact the network performances as the inter-arrivals and call durations. We compare these results to the classical distribution that is systematically considered in the models, the exponential law, and discuss its pertinence. For the space analysis, we establish a landscape of the usage of the Base Stations (BS). It has consisted in comparing the load for the different BS through Mexico city at different times and scales.

Beside, we propose a study of the handovers, *i.e.*, the fact, for a given user, to be bound to a new BS. As in many mobile traces, these handovers are only partially observable in our data set. Indeed, handovers do not appear explicitly in the logs, but are detected only when a call occurs on the new BS. Consequently, we are just able to determine if there is a BS change between two successive calls. This study relies on Palm Calculus [1, 2]. This theory gives practical tools to infer statistical properties of the handovers process from the process that describes the calls. The main results are: (i) a relevant estimator of the number of calls per time unit, (ii) a simple test on the independence between the two processes (calls and handovers), and (iii) an estimation of the handover distribution.

The paper is organized as follows. In Section 3, we describe our data set: available information, period of times, number of users, etc. In Section 4, we present the different results on the calls in time and space. Section 5 proposes a method to infer the statistical properties of the handovers, and presents the corresponding results. We conclude in Section 6.

2. Related work

The study of mobile phone data has been an active field during these last years. Plenty of topics has been covered such as mobile phone traffic, phone user ation, and human mobility [3]. In our study as in [4], we have the opportunity to analyze non-sparsified CDRs that represents a 1-year nationwide data set presented in [5].

The amount of mobile phone traffic has an overriding impact on the quality of service. The understanding of time evolution and spatial arrangement of the activity, studied in [6, 7], helps to enhance the network infrastructure and its capacity. As an example, the traffic analysis brings around a set of tools to detect specific local events and anomalies [8, 9] that commonly induce overload [10]. Predict and adapt protocols to respond to high activity periods is a subtantial benefit [11]. Event detection can also bring a criteria to identify user by his religion, soccer team, and music bands for example [12].

The traffic is the result of a causal chain where users and the way they communicate to each other are the starting point. From CDR, it is possible to understand better the human behavior and predict the traffic. For instance, [13] defines categories of mobile call profiles and classifies network usages accordingly and [14] makes the links between user phone usage and personnal behavior. To our knowledge, none of these studies consider the call duration as an information pool whereas its great impact on the load and on the intensity of social relations.

Complementarly, many studies focus on human mobility and tend to characterize, predict and model spatial individual mobility [7, 15, 16]. As user mobility seems to be unique [17], we can now predict next moves according to the mobility footprint. However, these works presume a precise knowledge of the switching from a spatial point to a new one. Whereas, in many CDRs, the handover times are unknown, we suggest in this paper that handover times and call events are pretty dependant.

3. Data set description

For this analysis, we use a CDR data set from a major mobile operator in Mexico. This CDR trace contains one year of geolocalized phone calls all over the country of Mexico. The dataset is anonymised. For each phone call, we have the timestamp in second, a phone Id of the subscriber originating the call, the phone Id of the user receiving the call, the call duration in second, and the BS of the telco company that routed the call (incoming or outgoing). For 77% of call records, there is one location which determines the location of the phone user belonging to the telco company (either the callee or the caller). If both caller and callee are clients of the telco company, then the location of the call record is randomly assigned to one of them. The trace is starting from the January 1, 2014 and ending on the December 31, 2014. It contains the whole 2014 year. For this period, we have more than 4.75 billions of calls. These geolocalized calls represent around 6% of global internal calls in the country of Mexico. As in our study we focus on the handovers, we will mostly consider the geolocalized calls. We can notice on figure 1, the missing locations are uniformly distributed and so will not have any impact on our results because it only decreases the ratio of calls we consider. This subset of calls is representing the activity of 7,700,208 telco users during one year.

The activity varies through time at several scales. During the day (from midday to

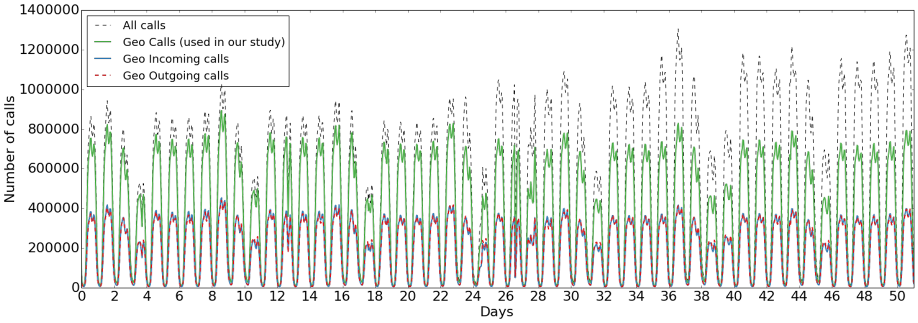


Figure 1: Mean number of calls for 51 days. The activity for a 51-day period extracted from our data set that corresponds to the 6th February to the 1st April 2014. It shows the variations of activity for each hour and for four different sets of calls : all the calls, geolocalized calls, outgoing geolocalized calls and incoming geolocalized calls. One can observe that there are the same number of incoming and outgoing calls. Geolocalized calls represents 76% all of the calls. We will use this data for the experiments that follow.

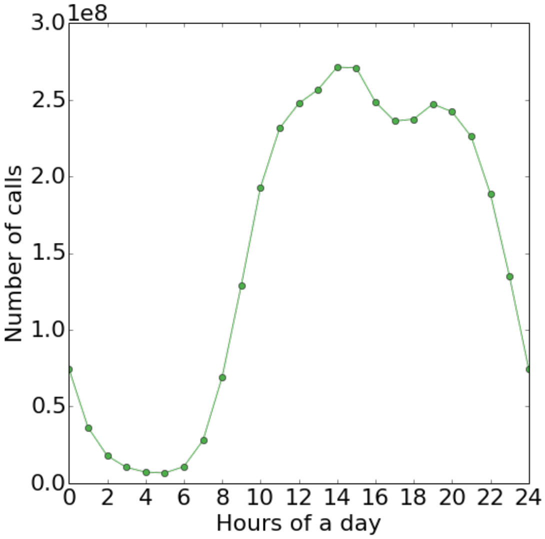


Figure 2: The number of calls during the 1-year period as function of the hours of the day. We note a period of lower activity during the night and higher activity during the day. The peak is reached at 2pm.

8pm), the activity is greater than during the night. The number of calls as function of the hours of the day (Figure 2) points out the typical period of lower activity during the night and greater activity during the day. Although the number of calls varies during the day, it also varies between different days. For instance, the activity during weekdays is greater than during the week-end. The peak is reached on Friday at 6 pm just after the end of the work.

There is a daily cyclo-stationarity in our signal. People are organized on a daily base of 24 hours such that the activity signal will have statistical properties that vary cyclically with time and can be viewed as multiple interleaved stationary processes. To show this intuitive point, an Empirical Mode Decomposition (EMD) [18] is performed on the activity signal, the number of calls per hour during 51 days. The EMD allows to represent the non-stationary signal as sum of zero-means Intrinsic Mode Function (IMF) and one residue. Figure 3 gives the decomposition of the global call activity in high and low frequencies. The IMF 2 to 5 clearly gives a daily periodic signal (a spectral analysis also gives an harmonic decomposition in days of the signal) which validate the cyclo-stationarity of the activity signal and the fact that globally, people are used to call or not at the same moment of the day. The high frequency IMF 1 is also plotted on Figure 3. We plot in red the mean of the residual. The signal is clearly oscillating around the mean in a compact envelope with few extra peaks of activity. The low frequency signal is useful when one tends to detect special events and anomalies on the activity.

In the mobile trace, the usage change drastically from a user to another as depicted on Figure 4. 75% of users have less than 2 calls per day whereas 25% have more than 10 calls.

In a mobile data trace, a lot of measures are quite heterogeneous like the number of contacts, the number of calls and the time between two calls. We show the distribution of the number of calls per day (Figure 4). We note that around 25% of the users have more than two calls per day whereas 25% of users have less than 10 calls per week. Running a handover study on a very long period of time will not make sense in such condition of strong heterogeneity. Indeed, it is impossible to determine rather if the user change precisely his location during a long inactive period. We do need a weak

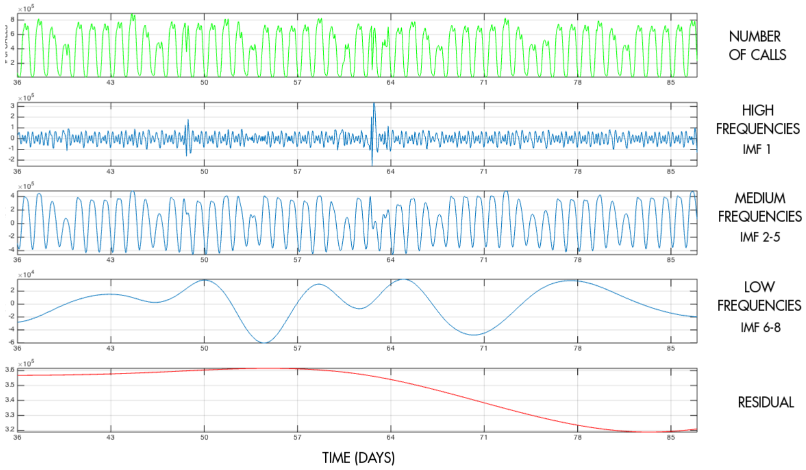


Figure 3: EMD of the signal linked to the number of calls per hour. From top to bottom, there is the original signal, high to low frequencies. One can clearly identify a day oscillation in the IMF 2-5. IMF 1 is high frequency variation and other IMF (6 to 8) are low frequencies.

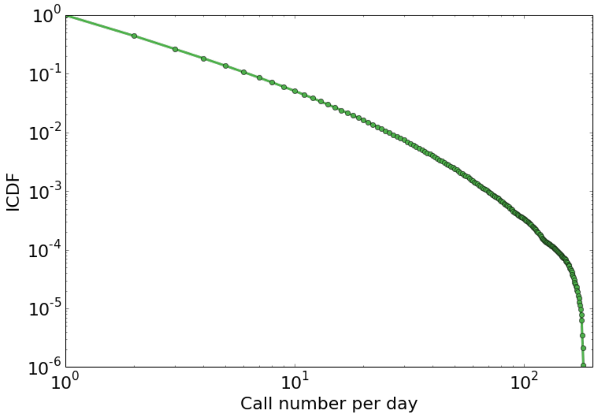


Figure 4: ICDF of the number of calls per user. The activity of users is heterogeneous, many people have few calls and some others have an active usage of the voice channel.

hypothesis on the stationarity of the signal. As we want to catch the movements of people during the day, we decided to cut all the signal (one year long) by slots of 2 hours. During each 2-hour period, we consider that the signal is stationary.

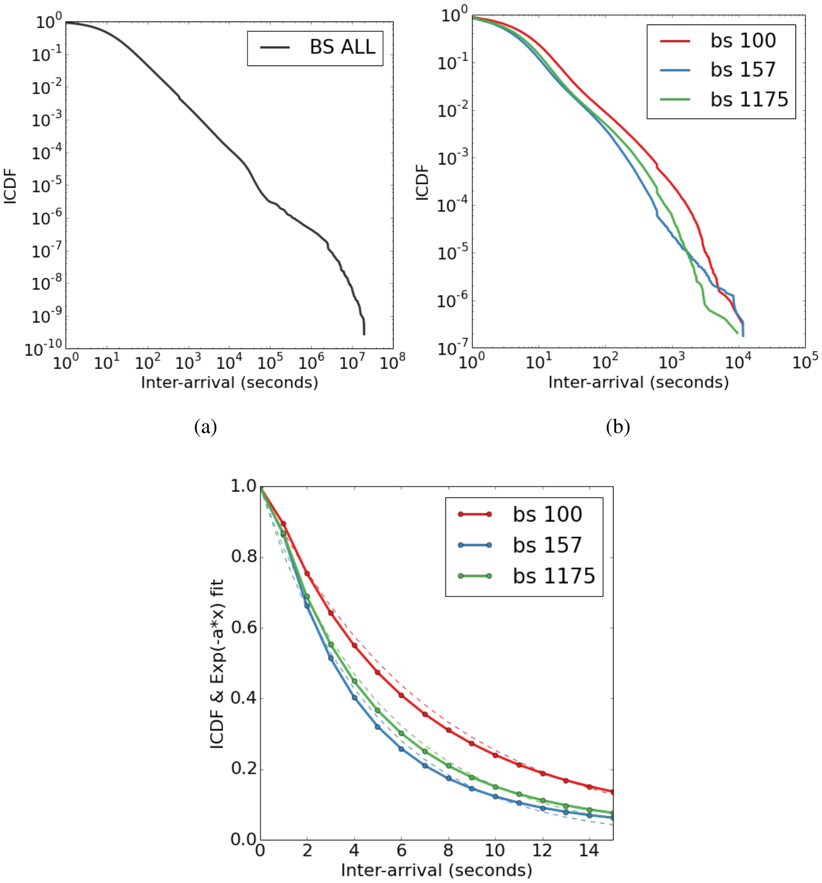
4. Call Analysis

In the two next sections, we analyze inter-arrival times between calls and call durations. These two quantities are the main input of queuing theory. The inter-arrivals describe the traffic nature, *i.e.*, the distribution of the clients arriving in the queue. The duration of a call is related to the service time of a client once it accesses to a resource. In our context, a resource is a couple slot-frequency or a set of resource blocks depending on the generation of cellular network we consider. In most of the queuing models, both inter-arrivals and call durations are supposed to be independently and exponentially distributed, leading to the famous $M/M/.$ queues. The reader can refer to [19], for a deeper presentation of queuing models applied to cellular networks. This assumption on the exponential distribution is common when considering phones traffic and call durations [20]. For the call duration, it is the distribution tail that is supposed to be exponential. Indeed, the first interval of the distribution is known as non exponential, because call durations cannot be less than a few seconds. But, the exponential assumption still offers a good approach as it is the tail distribution, "the big clients", that impacts the performance of the system.

This part of the analysis aims to study the statistical nature of the traffic, and to verify if the exponential assumption still holds.

4.1. Inter-arrivals on a Base station

When a user is calling someone, the origin and the destination are linked to a single BS. The attached BS are the first and last steps of the routing. In the trace, we only have one location which is the coordinates of the attached BS of the origin or the destination. Even if we miss many calls and so the activity of a BS is underestimated by a factor around 10, the distribution may be the same. The inter-arrival between two calls is overestimated.



(a)

(b)

(c)

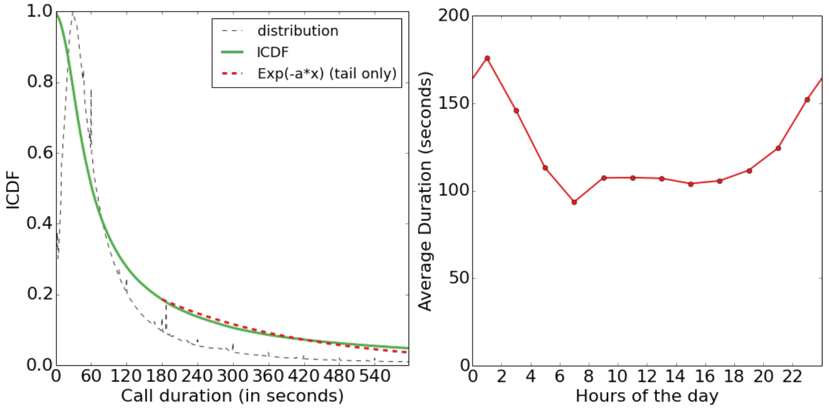
Figure 5: (a) For each BS, inter-arrival between two consecutive calls is computed and add to the global distribution. The plot is obtained by merging all the distributions. (b) For 3 specific BS, that corresponds to the 40%, 30% and 10% more active BS (60%, 70%, and 90% in terms of load) the distribution of the inter-arrival time in second between two consecutive calls is plotted in log-log scale. (c) For the same 3 specific BS, the ICDF from 0 to 15 seconds is fitted by an exponential function (dashed lines). For practical reason, x-axis is shifted by 1 second, we can so take the log as all values are strictly positive.

In Figure 5a, we plot the distribution, more precisely, the inverse cumulative distribution function, of the inter-arrivals. It corresponds to the time between two successive calls to a same BS. The distribution at the network scale, that gathers all geolocalized calls, is plotted in Figure 5a. It shows that the inter-arrivals range from 0 to several hours. These very high values of inter-arrivals could correspond to periods where a BS is switched off (for maintenance or other reasons). Also, the figure shows that 99% of the samples are less than 180 seconds, and 80% less than 21 seconds. The fact to consider all samples lead thus to a very large range of values, with a high proportion of samples with small inter-arrivals corresponding to peaks of traffic during the day, great inter-arrivals corresponding to the night traffic, and even inter-arrival of several hours. Also, we perform the same statistic evaluation for specific BS and time ranges. Indeed, these statistics help to dimension the network, which is usually performed with regard to the peak of traffic. We are thus interested to the traffic nature when the network is loaded. We considered three particular BS at the peak of traffic. We have first ordered all the BS as function of their load and choose three BS (BS numbered 1175, 157 and 100) that are respectively at 60%, 70%, and 90% in this classification. The distributions are shown in Figure 5b. For these distributions, at least 80% of the samples are less than 15 seconds (12 times less than the case with all samples). The three distributions have been fitted with an exponential law, represented by the dotted lines in Figure 5c. Even if it does not match exactly, the exponential is obviously very close to these distributions. The parameters of the exponential are 0.14, 0.19, and 0.21 and correspond to the mean number of calls per second. The standard deviation errors of the fit is respectively 0.0005, 0.0009 and 0.0007. The assumption on Poisson traffic is thus verified in our case.

4.2. Call duration

Here, we propose a study on the duration of a call. For each call for which the destination replied, there is a duration in second. The duration of a call is one of the parameter that has a major impact on the load.

We plot this distribution from our trace, by extracting a single duration of a random call per user. Each user counts only for one in the distribution 6a. In our trace, a



(a)

(b)

Figure 6: (a) ICDF of call duration in second (green), distribution of call duration in second normalized by the maximum to fit in the plot (black dashed line) and fit of the ICDF tail by exp^{-ax} (red); (b) Average for the whole year of duration calls for each 2 hours slot

long call is cut in several 10-minute calls. So, the distribution is ending at 10 minutes because the end of the tail is unknown. Apart from that, the ratio of long calls is quite small and the 10-minute sessions have a very small impact on the average and quartile results. We also noticed that there are more values when the number of seconds corresponds to a minute like 60s, 120s,... It is probably due to external artifacts like per-minute billing. The peak is reached for 34 seconds. The average duration of a call is 121 seconds and 25% of calls last more than 30 seconds whereas 75% last less than 2 minutes. All in all, 50% of calls take between 30 seconds and 2 minutes. The fit of the ICDF tail by an exponential function $x \mapsto exp(-a * x)$ gives $a = 0.004$ with $perror = 2.5 * 10^{-5}$. The behavior of the tail still tends to be very close to an exponential function as many studies already noticed it. This long tail induces an heterogeneity for the duration parameter, many durations are around 30 seconds and 2 minutes but some calls are still quite long.

In figure 6b, the time is divided in 12 slots of 2 hours each and the average duration is computed. From 6am-8am to 0am-2am, the average of the call duration is increasing. As the day is going, people tends to exchange more during a voice communication.

Then during the night (2am to 6am) people who answer do not take the time for long conversations. The shortest durations are recorded between 6am and 8am. According to parts of the day, the duration is changing. The average can double from a slot time to an other. This preliminary study on duration points out the fact that duration is not stationary and homogeneous but contains a lot of information that is useful to refine models or adapt performance of telco companies. These starting observations may help to refine models and improve performance.

5. Handover analysis

Data collected describes sent and received calls of users. For each call, the localization of the BS associated to the user is known. It allows us to know the BS location at the time the calls are made. Based on this knowledge, we can study the statistical properties of the BS changes, *i.e.* the different times at which a user is associated to a new BS. It reflects a certain vision of the users mobility and may be interesting for the telecoms operator as it corresponds to handovers that it has to manage. But these times are only partially observable: we are able to detect that between two successive calls the user is not bound to the same BS but we do not know when it does happen exactly between these two calls.

In this Section, we propose two estimators. The first one describes the mean number of handovers per time unit, and the second one is related to the cumulative distribution function (CDF) of the time between handovers. Also, we propose a simple test that allows us to check if the two processes, calls and handovers, are dependent. The different computations and proofs rely on Palm calculus. This mathematical framework offers a set of tools on stationary point processes. The reader can refer to [1] for the definition and tools of Palm Calculus in $I\!R$, or [2] for a more pedagogic introduction and its application in $I\!R^2$. As it will be shown, Palm calculus is particularly adapted to this study.

A stochastic point process is a random variable. It can be seen as an ordered set of points distributed in $\mathbb{R}$. The observation of a set of events occurring at different times can thus be modeled through a stochastic point process. Therefore, calls and handovers

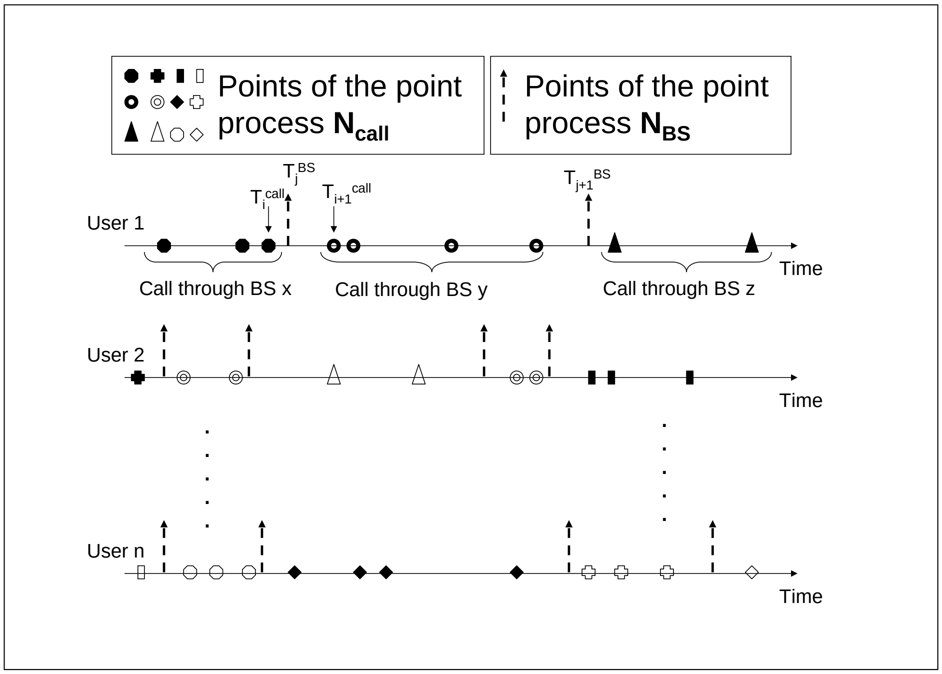


Figure 7: Description of the two-point processes N_{call} and N_{BS}. The point process N_{BS} is unobservable but the change of marks/BSs at the time calls (at points of N_{call}) allows us to know the intervals of N_{call} where they are located and to derive statistical properties of N_{BS}.

can be modeled through two-point processes. They are represented in Figure 7. The first point process is denoted N_{call}. A sample represents the time of the calls for a user. At the time of a call, we know the BS which the user is bound. Formally, it can be seen as a mark associated to the point process N_{call}. In Figure 7, we used different patterns to represent the points of N_{call} and its associated marks: a given pattern corresponding to a given BS/mark. For instance, when the user 1 is bound to BS x, the points/calls are depicted through black discs. When the user is bound to BS y, it is black ring, etc. A handover of user is thus detected when the mark of N_{call} changes. This marked point process is an exact representation/model of the data set in our possession.

The second process is N_{BS} and is depicted through the vertical arrows in the figure. It represents the changes of BS, the handovers, of a user. Our data set does not describe N_{BS}, but the marked process N_{call} allows us to determine between which calls there was a BS change, or equivalently between which points of N_{call} there is a point of

N_{BS}. For instance, for user 1 in Figure 7, we observe a change of BS, from BS x to BS y, between the points/calls T_i^{call} and T_{i+1}^{call}. Consequently, we infer the presence of a point of N_{BS} between the points T_i^{call} and T_{i+1}^{call}.

Formally, N_{BS} and N_{call} are random variables taking their values in the counting measures set on $(\mathbb{R}, \mathcal{B})$ (where $\mathcal{B}$ denotes the Borel σ-field of $\mathbb{R}$). We will use this definition in the different formulas, but as previously mentioned, it is more convenient to see a sample as a set of points (the support of the counting measure). A sample of N_{call} and N_{BS} can thus be seen as a set of points in $\mathbb{R}$, and correspond to the different time calls (N_{call}) and BS changes (N_{BS}) for a given user (a sample = a user).

A rapid analysis of the data showed that the process N_{call} is not ergodic, *i.e.*, statistics made on a given sample does not allow to obtain convergent estimators. For instance, the mean number of calls per time unit are very different from a user to another. The different statistical estimators that are derived in this section are then systematically based on all samples/users. In other words, we do not make statistics as the average of the observable quantities on large period of times, but instead we consider an event for each user/sample, the time between two calls for instance, and we compute the average of this event over all users/samples. We assume that the two-point processes are stationary. From the statistical point of view, we assume that the process is stationary on the interval of times where the statistics are computed. In the numerical results, the statistics are then given for different periods in the day. We also assume that there is at most one point of N_{BS} in an interval of N_{call}. It is a simplification of the reality, as the observation of a BS change may be, in practice, composed of a set of handovers. The impact of this assumption on the results are discussed in the next paragraphs when presenting the results.

5.1. Intensity

The first quantity that is studied is the intensity of N_{BS}, denoted λ_{BS}, *i.e.* the mean number of BS changes per unit time. We propose an estimator $\widehat{\lambda_{BS}}$ of this quantity. Let Ω be the set of samples (our data set). The samples in Ω are assumed to be independent.

Our estimator is obtained through the application of Palm calculus. The points of N_{call} (respectively N_{BS}) are denoted $(T_i^{call})_{i\in\mathbb{Z}}$ (respectively $(T_i^{BS})_{i\in\mathbb{Z}}$), in ascending order, and where $[T_0^{call}, T_1^{call}]$ (respectively $[T_0^{BS}, T_1^{BS}]$) is the interval that contains the origin. We apply the Neveu's exchange formula ([1] page 21) to the two-point processes N_{BS} and N_{call} for a function $f = 1$. We obtain:

$$\lambda_{call} = \lambda_{BS}\mathbb{E}^0_{N_{BS}}\left[\int_0^{T_1^{BS}} N_{call}(dx)\right] \tag{1}$$

$\mathbb{E}^0_{N_{BS}}[.]$ is the Palm expectation with regard to the process N_{BS}. Palm expectation, or Palm measure, may be seen as the probability measure under the condition that there is a point of the point process at the origin. The point process indexed under the expectation notation $\mathbb{E}^0_{N_{BS}}$ (N_{BS} here) indicates which point process is supposed to have a point at the origin. It is worth noting that quantities under the classical and Palm expectation lead to different values. For instance, $\mathbb{E}\left[T_1^{BS}\right]$ and $\mathbb{E}^0_{N_{BS}}\left[T_1^{BS}\right]$ differs. $\mathbb{E}\left[T_1^{BS}\right]$ is the time from the origin (an arbitrary time) to the next BS change. It is thus the residual time to the next BS change. $\mathbb{E}^0_{N_{BS}}\left[T_1^{BS}\right]$ is the time between two BS changes. Indeed, under Palm expectation we know that there is a point of N_{BS} at 0 and we evaluate the time to the next BS change T_1^{BS}.

In Equation (1), remind that $N_{call}(.)$ is a counting measure, and $\int_0^{T_1^{BS}} N_{call}(dx)$ is thus equal to $N_{call}([0, T_1^{BS}])$. Under Palm expectation, $\int_0^{T_1^{BS}} N_{call}(dx)$ is thus the mean number of points of N_{call} between two successive points of N_{BS}. Consequently, this quantity can be fully determined/estimated based on our samples as we do not need the exact location of the points of N_{BS}. For instance, in Figure 7, a sample of this quantity (for user 1) is the number of points of N_{call} between points T_j^{BS} and T_{j+1}^{BS} (equals to 4). The estimator is then:

$$\widehat{\lambda_{BS}} = \frac{\widehat{\lambda_{call}} card(\Omega)}{\sum_{\mathcal{N}_{call}\in\Omega} \mathcal{N}_{call}([T_i^{BS}, T_{i+1}^{BS}])} \tag{2}$$

where $\widehat{\lambda_{call}}$ is an estimator of λ_{call}.

Equation (2) corresponds exactly to equation 1, but wrote in a simpler form. Here, we pick one interval of N_{BS} for each sample. The value of i does not matter and may be different from one sample to another. Due to the stationarity constraint, we divide times of the day to slots of 2 hours. The estimation of λ_{BS} is then performed independently for each slot. We take one sample for each user and each day. Results are shown in Table 1. The number of considered samples is shown in the last column of the table. With the constraint of 2 hours, all samples are not taken into account. Indeed, we consider only samples with at least two handovers/movements, otherwise it is obviously impossible to apply the method. The results show that the number of handovers stays more or less constant during all days. With the filter that we apply on the data set, the results tend to show that, in average, a user moves rarely more than two times on these slots. Clearly, our method leads to an over estimation of the real intensity λ_{BS}. But, a classical method consisting in evaluating the time between two handovers with the same constraint on the stationarity should lead exactly to the same problem. Moreover, in our study, we do not have the exact time between two handovers, such an approach would consequently be impossible.

Hours	λ_{BS} (second)	λ_{BS} (hour)	Omega
0-2AM	0.00058	2.09	1 044 566
2-4AM	0.00061	2.19	265 693
4-6AM	0.00061	2.18	129 196
6-8AM	0.00060	2.17	215 682
8-10AM	0.00058	2.07	1 676 401
10-12AM	0.00059	2.11	6 899 027
12-14PM	0.00059	2.14	9 543 073
14-16PM	0.00058	2.10	10 545 188
16-18PM	0.00057	2.07	8 899 166
18-20PM	0.00058	2.07	8 409 011
20-22PM	0.00056	2.01	7 574 970
22-24PM	0.00056	2.01	4 058 205

Table 1: Results for the estimation of λ_{BS}.

5.2. Dependency test

An important assumption to estimate the distribution of the time between two successive points of N_{BS} is the dependency between the two processes N_{BS} and N_{call}. A formal hypothesis test is impossible to perform as N_{BS} is not fully observable. Therefore, we propose a simple test, based on the length of the intervals $[T_i^{call}, T_{i+1}^{call}]$ where the points of N_{BS} are located, to infer the dependency between the two processes.

According to Palm Calculus, if we pick a point X in $I\!R$ independently of a stationary point process, e.g. N_{call}, this point will be likely located in a "big interval". More precisely, the mean of the interval length $[T_i^{call}, T_{i+1}^{call}]$ where X is located will be greater than the mean size of the interval of the point process ($\frac{1}{\lambda_{call}}$ here). Intuitively, as "big intervals" occupy more space, X is likely located in one of them. For instance, with a Poisson point process the mean interval length where X is located is two times greater than the other intervals (in average). It is the famous Feller paradox ([1] pages 33 and 295). As the process is stationary, pick a random point X or a fix point leads to the same results. By convenience we consider the origin. The interval where the origin is located is $[T_0^{call}, T_1^{call}]$, and its mean length is equal to $\mathbb{E}[T_1^{call} - T_0^{call}] = 2 \cdot \mathbb{E}[T_1^{call}]$ (consequence of the stationarity of the point process). The mean length of this interval depends on the distribution of the process, but it can be easily calculated with the Palm inversion formula ([1] page 20). We give below the computation details but it is a classical result of Palm calculus (see [21] for instance where it is applied to a mobility study). In the first equation below, θ_t is the shift operator. Here, it shifts the points of N_{call} of a time t (meaning that $N \circ \theta_x(C) = N(C - t)$ for an interval C in $\mathbb{R}$, or more formally C in $\mathcal{B}$). We get:

$$\begin{aligned}
\mathbb{E}\left[T_1^{call}\right] &= \lambda_{call}\mathbb{E}_{call}^{0}\left[\int_0^{T_1^{call}} T_1^{call} \circ \theta_t dt\right] && (3)\\
&= \lambda_{call}\mathbb{E}_{N_{call}}^{0}\left[\int_0^{T_1^{call}} (T_1^{call} - t) dt\right] && (4)\\
&= \frac{\lambda_{call}}{2}\mathbb{E}_{N_{call}}^{0}\left[\left(T_1^{call}\right)^2\right] && (5)\\
& && (6)
\end{aligned}$$

If N_{BS} is independent of N_{call}, a point of N_{BS} behaves as the random point X presented earlier or the origin. Therefore, if the two processes are independent the mean interval lengths (of N_{call}) where the points of N_{BS} are located must equal to

Equation (5). If they are different, it proves that the two processes are dependent. Unfortunately, it does not prove the independence in case of equality. It is worth noting that these two quantities do not depend on the exact locations of the points of N_{BS} but only on the interval lengths of N_{call} available from our data set.

The results are shown in Table 2. Before describing the results, we give some elements on the method we followed. We considered intervals of two hours during the day to obtain intervals where the two-point processes are assumed stationary. Each temporal window was processed independently. For each user, we draw randomly one of the handovers and we measured the interval $[T_i^{call}, T_{i+1}^{call}]$ where it lied. The result is the column "Handover interval" in the table. Beside, we selected an interval $[T_i^{call}, T_{i+1}^{call}]$ randomly chosen for each user and estimate these two first moments (compute as the average over all users). It leads to estimators of $E^0_{N_{call}}[T_1^{call}]$ and $E^0_{N_{call}}[(T_1^{call})^2]$, from which we deduce $E[T_1^{call} - T_0^{call}]$ (equal to two times Equation 5).

In Table 2, we can observe, as expected, that handover happens in interval with a greater length in average with regard to $E^0_{N_{call}}[T_1^{call}]$. But, their mean lengths should equal to the 4^{th} column. We can observe a difference of approximately 30% between these two quantities. With the number of samples used in the different computations, that are given in the last columns, the confidence intervals are close to 0 for all these estimators, and so does not explain the gap. This difference seems to show that the two processes are correlated. A possible interpretation of this phenomena, is that mobile users may call before a departure, at their arrival, or during the path, and consequently are likely to call when they are in movement or just after/before a movement. This result may present a bias as we do not know the number of handovers between two calls. Indeed, several handovers may happen between two successive calls. Therefore, our choice of the intervals with an handover would be different if the number of handovers is very different from an interval to another. Intuitively, in this case, we should more likely choose an interval with a great number of handovers than an interval with only a small one.

Hours	$E^0_{N_{call}}[T_1^{call}]$	Handover interval	$E[T_1^{call} - T_0^{call}]$	number of samples
0-2AM	591.79	942.25	1324.77	35058
2-4AM	578.03	923.05	1377.29	9280
4-6AM	564.49	1007.60	1453.25	4531
6-8AM	565.61	1073.24	1515.29	8445
8-10AM	634.54	1152.25	1632.98	62660
10-12AM	719.78	1252.82	1797.30	193054
12-14PM	727.06	1277.26	1825.02	237929
14-16PM	718.87	1275.12	1816.50	260823
16-18PM	716.41	1298.34	1827.37	218267
18-20PM	715.20	1264.40	1810.68	218915
20-22PM	687.99	1242.35	1741.00	200104
22-24PM	655.38	1128.19	1628.88	118539

Table 2: Results on the dependency test.

5.3. Distribution

In this section, we describe a method to obtain estimations of the distribution of N_BS. More precisely, we assess the cumulative distribution function (CDF) of T_1^{BS} under the classical probability measure ($\mathbb{P}\left(T_1^{BS} \leq x\right)$) and Palm measure ($\mathbb{P}^0_{N_{BS}}\left(T_1^{BS} \leq x\right)$). Under the Palm measure, it describes the distribution of the time between two successive handovers. Under the classical measure, it is the time to the next handover: given a user at an instant t, it is the time to the next handover.

We do know the intervals where the points of N_{BS} are distributed. In each of these intervals, we draw the point N_{BS} uniformly. It would correspond to the real distribution in case of independence of the two processes. But, as we have seen in the previous Section, independence does not hold here and our method is thus not exact.

From these samples we compute the empirical estimator of $\mathbb{P}\left(T_1^{BS} < u\right)$.

We detail below the method. A set of examples is given in Figures 8 and 9. The method:

- We set a common time t as the origin for all our samples. It is chosen arbitrarily

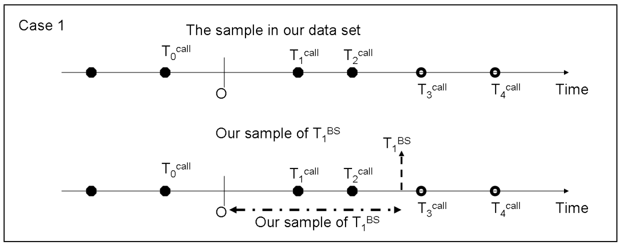


Figure 8: Case 1: the first point of T_1^{BS} is in the interval $[T_2^{call}, T_3^{call}]$. The sample of T_1^{BS} is then uniformly distributed in this interval.

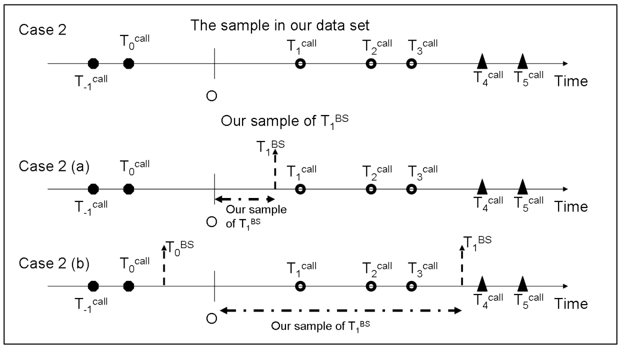


Figure 9: Case 2: there is a point of N^{BS} in the interval $[T_0^{call}, T_1^{call}]$. We draw a point of N_{BS} uniformly in this interval. It leads to two sub-cases: (Case 2(a)) if the point is in $[O, T_1^{call}]$, then we consider it as our sample of T_1^{BS}, (Case 2(b)) if the point is in $[T_0^{call}, O]$ then it corresponds to T_0^{BS}, so we look for the next interval that hosts a point of N_{BS} ($[T_3^{call}, T_4^{call}]$ in the figure) and we distribute uniformly our sample of T_1^{BS} in this interval.

and independently of the two processes. It is denoted O in the figures.

- To collect samples of points T_1^{BS}, we proceed as follows for each sample/user:
  - If the interval of N_{call} that contains the first point of N_{BS} is $[T_i^{call}, T_{i+1}^{call}]$ with $i > 0$, then we draw our sample uniformly in this interval. This case is illustrated in Figure 8 (Case 1).
  - If the interval of N_{call} that contains the origin, $[T_0^{call}, T_1^{call}]$, hosts a point of N_{BS}, then we draw a point uniformly in $[T_0^{call}, T_1^{call}]$. If it belongs to $[0, T_1^{call}]$ then we select this point as our sample (Figure 9 - Case 2(a)). Otherwise, the point that is obtained is T_0^{BS} and not T_1^{BS}. Consequently,

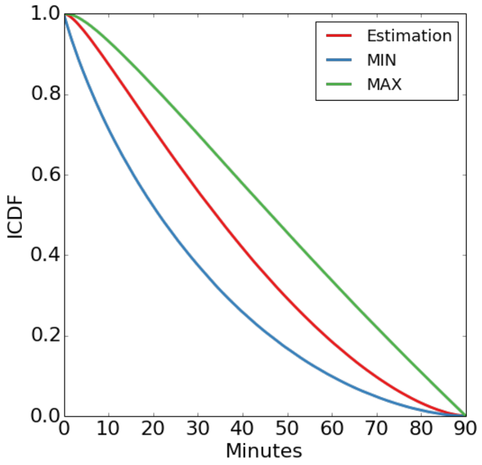


Figure 10: ICDF of T_1^{BS}.

we consider the next interval of N_{call} ($[T_i^{call}, T_{i+1}^{call}]$ with $i > 0$) that contains a point of N_{BS}. Our sample is then the point uniformly distributed in this interval (Figure 9 - Case 2(b)).

- From the collected samples, we calculate the empirical distribution of T_1^{BS} *i.e.* $\mathbb{P}\left(T_1^{BS} \leq x\right)$.

We also consider a lower and upper bound on the values of the samples that allows to bound the real distribution of T_1^{BS}. For the lower bound, we consider for each sample the beginning of the interval $[T_i^{call}, T_{i+1}^{c}all]$, thus T_i^{call}. For the upper bound we consider T_{i+1}^{call}.

The inverse cumulative distribution function (ICDF) is shown in Figure 10. The ICDF under the classical expectation, shows that handovers occurs between 0 and approximately 5400 seconds. The empirical distribution is close to a uniform distribution which would be a straight line between 1 (at 0) and 0 (at 5400 seconds). The two bounds do not present negligible differences with the approximated distribution. It can reach up a difference of 0.2 for the lower bound, and 0.15 for the upper bound.

6. Conclusion

This paper presented an analysis of calls in a cellular network from a CDR trace. In the first part, we assess the statistical properties of these calls. We exhibit a cyclo-stationarity of the number of calls per hour, with a lightweight different behavior in the week-end. Also, the distribution obtained for call durations and inter-arrivals have shown that the classical exponential still fit to the empirical one. It confirms the classical assumptions on phone traffic. Moreover, our study gives example of current loads observed in cellular networks that can be considered as input in queuing models.

In the second part, we have proposed a method to study user movements using Palm calculus. This theory offers a formal mathematical framework to obtain estimator on user movements. Consequently, we have proposed methods to estimate the intensity of user movements, a dependency test that allowed us to check if calls and movements are correlated, and a method to generate samples of user movements. A required property to apply this theory is that the considered processes must be stationary. As it is clearly not the case for our data set, we had to consider range of 2 hours. It led to a proportion of samples with no movements that could not be taken into account, and thus an over-estimation of user movements. Also, for moving users, the proposed dependency test seems to show that their movements are correlated to their calls.

Results of our study may be used in different ways. It can help to consider practical parameters in simulations and models. Results on the dependency between calls and movements still need to be improved. A more detailed characterization of this dependency could help to propose models able to generate joint calls and movements distribution. Also, we point out that our results on the call durations contain a lot of information by its variability through time and users. This quantity can help to improve models that describe social relationship between users. Taking advantage of this parameter can lead to identify people, define contacts between phone users, detect communities or predict links. A more fundamental work could consist in extending this study to non-stationary point process. The question is: may we rely on the cyclo-stationarity of the processes to derive equivalent estimators from Palm Calculus but applied to the full periods (complete weeks, months, or year).

References

[1] F. Baccelli, P. Brémaud, Elements of queueing theory : Palm-martingale calculus and stochastic recurrences, 2nd Edition, Springer, Berlin; New York, c2003, (TIT) Palm-martingale calculus and stochastic recurrences.

[2] D. Stoyan, W. S. Kendall, J. Mecke, Stochastic geometry and its applications, Wiley series in probability and mathematical statisitics, Wiley, Chichester, W. Sussex, New York, 1987, rev. translation of: Stochastische Geometrie.

[3] F. Calabrese, L. Ferrari, V. D. Blondel, Urban sensing using mobile phone network data: a survey of research, ACM Computing Surveys (CSUR) 47 (2) (2014) 25.

[4] P. Zerfos, X. Meng, S. H. Wong, V. Samanta, S. Lu, A study of the short message service of a nationwide cellular network, in: Proceedings of the 6th ACM SIGCOMM conference on Internet measurement, ACM, 2006, pp. 263–268.

[5] C. Sarraute, P. Blanc, J. Burroni, A study of age and gender seen through mobile phone usage patterns in mexico, in: Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on, 2014, pp. 836–843. doi:10.1109/ASONAM.2014.6921683.

[6] Y. Leo, C. Sarraute, A. Busson, E. Fleury, Taking benefit from the user density in large cities for delivering sms, in: Proceedings of the 12th ACM Symposium on Performance Evaluation of Wireless Ad Hoc, Sensor, &#38; Ubiquitous Networks, PE-WASUN '15, ACM, New York, NY, USA, 2015, pp. 55–61. doi:10.1145/2810379.2810393.

[7] M. C. Gonzalez, C. A. Hidalgo, A.-L. Barabasi, Understanding individual human mobility patterns, Nature 453 (7196) (2008) 779–782.

[8] Y. Dong, F. Pinelli, Y. Gkoufas, Z. Nabi, F. Calabrese, N. V. Chawla, Inferring unusual crowd events from mobile phone call detail records, in: Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD

2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part II, 2015, pp. 474–492.

[9] A. Dobra, N. E. Williams, N. Eagle, Spatiotemporal Detection of Unusual Human Population Behavior Using Mobile Phone Data, PLOS ONE 10 (3) (2015) e0120449+. arXiv:1411.6179, doi:10.1371/journal.pone.0120449.
URL `http://dx.doi.org/10.1371/journal.pone.0120449`

[10] U. Paul, A. Subramanian, M. Buddhikot, S. Das, Understanding traffic dynamics in cellular data networks, in: INFOCOM, 2011 Proceedings IEEE, 2011, pp. 882–890. doi:10.1109/INFCOM.2011.5935313.

[11] G. Heine, M. Horrer, GSM networks: protocols, terminology, and implementation, Artech House, Inc., 1999.

[12] P. Paraskevopoulos, T.-C. Dinh, Z. Dashdorj, T. Palpanas, L. Serafini, Identification and characterization of human behavior patterns from mobile phone data, Proc. of NetMob.

[13] D. Naboulsi, R. Stanica, M. Fiore, Classifying call profiles in large-scale mobile traffic datasets, in: INFOCOM, 2014 Proceedings IEEE, 2014, pp. 1806–1814. doi:10.1109/INFOCOM.2014.6848119.

[14] Y.-A. de Montjoye, J. Quoidbach, F. Robic, A. S. Pentland, Predicting personality using novel mobile phone-based metrics, in: Social computing, behavioral-cultural modeling and prediction, Springer, 2013, pp. 48–55.

[15] C. Song, T. Koren, P. Wang, A.-L. Barabási, Modelling the scaling properties of human mobility, Nature Physics 6 (10) (2010) 818–823.

[16] J. Candia, M. C. Gonzalez, P. Wang, T. Schoenharl, G. Madey, A.-L. Barabasi, Uncovering individual and collective human dynamics from mobile phone records, Journal of Physics A: Mathematical and Theoretical 41 (22) (2008) 224015.

[17] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, V. D. Blondel, Unique in the crowd: The privacy bounds of human mobility, Scientific reports 3.

[18] G. Rilling, P. Flandrin, P. Gonçalves, On empirical mode decomposition and its algorithms, in: Proceedings of IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing NSIP-03, Grado (Italy), 2003.

[19] L. Ponomarenko, C. S. Kim, A. Melikov, Performance Analysis and Optimization of Multi-Traffic on Communication Networks, 1st Edition, Springer-Verlag New York, Inc., New York, NY, USA, 2010.

[20] M. Zonoozi, P. Dassanayake, M. Faulkner, Mobility modelling and channel holding time distribution in cellular mobile communication systems, in: Global Telecommunications Conference, 1995. GLOBECOM '95., IEEE, Vol. 1, 1995, pp. 12–16 vol.1. doi:10.1109/GLOCOM.1995.500213.

[21] J.-Y. Le Boudec, Understanding the simulation of mobility models with palm calculus, Perform. Eval. 64 (2) (2007) 126–147. doi:10.1016/j.peva.2006.03.001.