

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

Yannick Léo, Anthony Busson, Carlos Sarraute, Eric Fleury

▶ **To cite this version:**

Yannick Léo, Anthony Busson, Carlos Sarraute, Eric Fleury. Call Detail Records to Characterize Usages and Mobility Events of Phone Users. ALGOTEL 2016 - 18èmes Rencontres Francophones sur les Aspects Algorithmiques des Télécommunications, May 2016, Bayonne, France. hal-01303707

HAL Id: hal-01303707

https://inria.hal.science/hal-01303707

Submitted on 18 Apr 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

Yannick Leo[14] and Anthony Busson[14] and Carlos Sarraute[2] and Eric Fleury[134]

[1] *ENS de Lyon, Lyon, France*
[2] *Grandata Labs, Buenos Aires, Argentina*
[3] *INRIA, France*
[4] *University of Lyon (UMR CNRS - ENS Lyon - UCB Lyon 1 - INRIA 5668)*

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 defined 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 dependent 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 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.

In this paper, we focus on the analysis of a trace represents 90 millions of Mexicans calling each others from the network/operator point of view. We 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.

Many studies focus on human mobility, we can now predict next moves according to the mobility footprint. Whereas, in many CDRs, the handover times are unknown, we suggest in this paper that handover times and call events are pretty dependent. Moreover, we propose a study of the handovers, *i.e.*, the fact, for a given user, to be bound to a new BS. We are just able to determine if there is a BS change between two successive calls. This study relies on Palm Calculus [BB03]. This theory gives practical tools to infer statistical properties of the handovers process from the process that describes the calls.

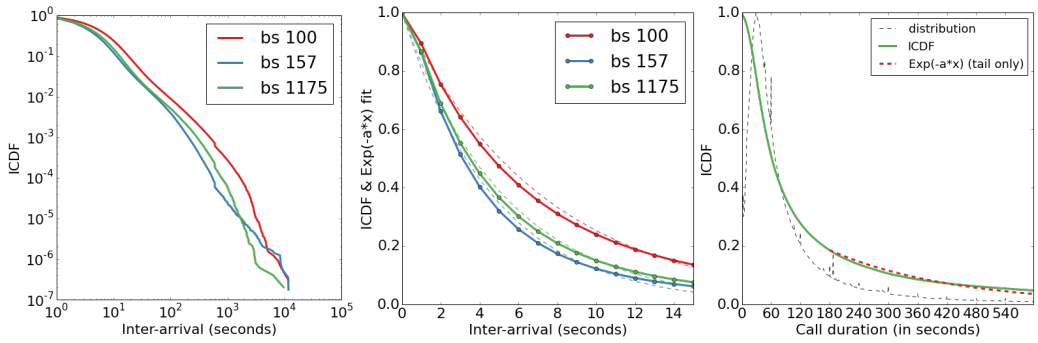
The main results are : (i) a relevant estimator of the number of calls per time unit, (ii) a simple test on the independence between the two processes (calls and handovers), and (iii) an estimation of the handover distribution.

2 Data set description

For this analysis, we use a CDR data set from a major mobile operator in Mexico [SBB14]. 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. 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 and the usage change drastically from a user to another, 75% of users have less than 2 calls per day whereas 25% have more than 10 calls.

We analyze inter-arrival times between calls and call durations. Whereas the inter-arrival times between calls seem to have a very long tail (Fig. 1a), if we only focus on the interval [0-15] seconds which represents 75% of the distribution. We show that these distribution are exponential distributions (Fig. 1b). The call duration tail seems to be also an exponential function $x \mapsto exp(-a*x)$ gives $a = 0.004$ with $perror = 2.5*10^{-5}$. 50% of calls take between 30 seconds and 2 minutes (Fig. 1c).



(a) (b) (c)

FIGURE 1: (a) 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. (b) 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. (c) 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).

3 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. Moreover, 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 [BB03] which 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. We introduce two stochastic point processes : calls (N_{call}) and handovers (N_{BS}). N_{call} denotes the time of the calls for a user. At the time of a call, we know the BS which the user is bound. N_{BS} 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}. Consequently, we infer the presence of a point of N_{BS} between some consecutives points T_i^{call} and T_{i+1}^{call}.

As the process N_{call} is not ergodic, we do not make statistics as the average of the observable quantities on large period of times, but instead we consider a single 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 2-hour interval 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.

3.1 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. After applying the Neveu's exchange formula ([BB03] page 21) to the two-point processes N_{BS} and N_{call} for a function $f = 1$. Under Palm expectation, we note $\mathcal{N}_{call}$ the mean number of points of N_{call} between two successive points of N_{BS}. The estimator $\widehat{\lambda_{BS}}$ is then :

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

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. In average, $\widehat{\lambda_{BS}} = 2.11$. Indeed, we consider only samples with at least two handovers/movements, otherwise it is obviously impossible to apply the method. 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 2-hour slots.

3.2 Dependency test

An interresting question 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.

$E^0_{N_{call}}[T_1^{call}]$	Handover interval	Independent Interval (theory)
656s	1153s	1646s

TABLE 1: Results on the dependency test.

According to Palm Calculus, if we pick a point X in $\mathbb{R}$ independently of a stationary point process, e.g. N_{call}, this point will be likely located in a "big interval". Even if we skip the technical part, we can note in

Table 1 that the handovers intervals (2^{nd} column) are quite bigger than the typical intervals $[T_i^{call}, T_{i+1}^{call}]$ (1^{st} column). Yet, it is still dependent, the difference between the handover interval size and the theoric size to be consider as independent is about 40%.

3.3 Distribution

In this section, we describe results obtained by a method to extimate 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 do not detail the method but it consists, for each sample, in picking a time T_1^{BS} uniformly between the first interval that contains a point of N_BS. 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}.

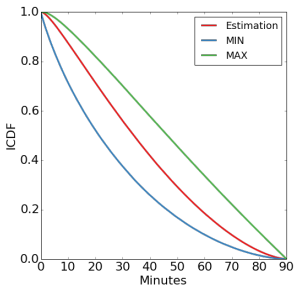


FIGURE 2: ICDF of T_1^{BS}. The inverse cumulative distribution function (ICDF) is shown in Figure 2. The ICDF under the classical expectation, shows that handovers occurs between 0 and approximately 5400 seconds. The empirical distribution is close to a uniform distribution which would be a straight 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.

4 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. 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. Also, for moving users, the proposed dependency test seems to show that their movements are correlated to their calls.

Références

[BB03] Francois Baccelli and Pierre Bremaud. *Elements of queueing theory : Palm-martingale calculus and stochastic recurrences*. Springer, Berlin ; New York, 2nd ed. edition, c2003. (TIT) Palm-martingale calculus and stochastic recurrences.

[SBB14] C. Sarraute, P. Blanc, and 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*, pages 836–843, Aug 2014.