Surprising Patterns for the Call Duration Distribution of Mobile Phone Users

Pedro O.S. Vaz de Melo[1,4], Leman Akoglu[2,3,4], Christos Faloutsos[2,3,4], and Antonio A.F. Loureiro[1]

[1] Universidade Federal de Minas Gerais
[2] Carnegie Mellon University
[3] SCS, School of Computer Science
[4] iLab, Heinz College
olmo@dcc.ufmg.br, lakoglu@cs.cmu.edu, christos@cs.cmu.edu, loureiro@dcc.ufmg.br

Abstract. How long are the phone calls of mobile users? What are the chances of a call to end, given its current duration? Here we answer these questions by studying the call duration distributions (CDDs) of individual users in large mobile networks. We analyzed a large, real network of *3.1 million* users and more than one *billion* phone call records from a private mobile phone company of a large city, spanning $0.1TB$. Our first contribution is the TLAC distribution to fit the CDD of each user; TLAC is the truncated version of so-called *log-logistic* distribution, a skewed, power-law-like distribution. We show that the TLAC is an excellent fit for the overwhelming majority of our users (more than 96% of them), much better than exponential or lognormal. Our second contribution is the *MetaDist* to model the collective behavior of the users given their CDDs. We show that the *MetaDist* distribution accurately and succinctly describes the calls duration behavior of users in large mobile networks. All of our methods are fast, and scale linearly with the number of customers.

1 Introduction

In the study of phone calls databases [18,20,17], a common technique to ease the analysis of the data is the summarization of the phone calls records into aggregated attributes [10], such as the aggregate calls duration or the total number of phone calls. By doing that, the size of the database can be reduced by orders of magnitudes, allowing the execution of most well known data mining algorithms in a feasible time. However, we believe that such representation veils relevant temporal information inherent in a user or in a relationship between two people. When all the information about the phone calls records of a user is aggregated into single summarized attributes, we do not know anymore how often this user calls or for how long he talks per phone call. One may suggest, for instance, to use descriptive statistics such as mean and variance to describe the duration of the user's phone calls, but it is well known that the distribution of these values is highly skewed [20], what invalidate the use of such statistics.

In this paper, we tackle the following problem. Given a very large amount of phone records, what is the best way to summarize the calling behavior of a user? In order

J.L. Balcázar et al. (Eds.): ECML PKDD 2010, Part III, LNAI 6323, pp. 354–369, 2010.

to answer this question, we examine phone call records obtained from the network of a large mobile operator of a large city. More specifically, we analyze the duration of hundreds of million calls and we propose the *Truncated Lazy Contractor* (TLAC) model to describe how long are the durations of the phone calls of a single user. Thus, the TLAC models the Calls Duration Distribution (CDD) of a user and is parsimonious, having only two parameters, the *efficiency* coefficient ρ and the *weakness* coefficient β. We show that the TLAC model was the best alternative to model the CDD of the users of our dataset, mainly because it has a heavier tail and head than the log-normal distribution, that is the most commonly used distribution to model CDDs [7].

We also suggest the use of the TLAC parameters as a better way to summarize the calls duration behavior of a user. We propose the *MetaDist* to model the population of users that have a determined calls duration behavior. The *MetaDist* is the meta-distribution of the ρ_i and β_i parameters of each user's i CDD and, when its isocontours are visualized, its shape is surprisingly similar to a bivariate Gaussian distribution. This fascinating regularity, observed in a significantly noisy data, makes the *MetaDist* a potential distribution to be explored in the direction of better understanding the call behavior of mobile users.

Thus, in summary, the main contributions of this paper are:

- The proposal of the TLAC model to represent the individual phone calls durations of mobile customers;
- The *MetaDist* to model the group call behavior of the mobile phone users;
- The use of the *MetaDist* and the *Focal Point* to describe the collective temporal evolution of large groups of customers;

As an additional contribution, we show the usefulness of the TLAC model. We show that it can spot anomalies and it can succinctly verify correlations (or lack thereof) between the TLAC parameters of the users and their total number of phone calls, aggregate duration and distinct patterns. We also emphasize that the TLAC model can be used to generate synthetic datasets and to significantly summarize a very large number of phone calls records.

The rest of the paper is organized as follows. In Section 2, we provide a brief survey of other work that analyzed mobile phone records. In Section 3, we describe our proposed TLAC model and we show its goodness of fit. The *MetaDist* and the analysis on the temporal evolution of the collective call behavior of the customers of our dataset is shown in Section 4. In Section 5, we discuss the possible applications for our results and, finally, we show the conclusions and future research directions in Section 6.

2 Related Work

A natural use for a mobile phone dataset is to construct the social network from its records [10,8]. In [16,17], the authors construct a network from mobile phone calls records and, from it, they make a detailed analysis of its network properties. They identified relationships between node weights and network topology, finding that the weak ties are commonly responsible for linking communities, thus having a high betweenness centrality or low link overlap. Moreover, in [8], the authors verified that the persistence

of an edge is highly correlated to its reciprocity and to the topological overlap and, in [4], the authors explore communication networks in order to verify the patterns that occurs in its cliques. It is also common to analyze the networks from mobile companies in order to improve their services. For instance, in [9,3], the authors proposed a framework and data structures for identifying fraudulent consumers on telecommunication networks based on their degree distribution and dynamics and, in [15], the authors proposed metrics that can be employed by a business strategy planner involved in the telecom domain.

Another use for a mobile phone dataset is to study the individual attributes of the users. In [18], the authors proposed the DPLN distribution to model the distributions of the number of phone calls per customer, the total talk minutes per customer and the distinct number of calling partners per customer. In [7], the authors analyzed mobile phone calls that arrived in a mobile switch center in a GSM system of Qingdao, China, and they found that the duration of the phone calls is best modeled by a log-normal distribution. However, in [20], the authors studied the duration of mobile calls arriving at a base station during different periods and found that they are neither exponentially nor log-normally distributed, possessing significant deviations that make them hard to model. They verified that about 10% of calls have a duration of around 27 seconds, that correspond to calls which the called mobile users did not answer and the calls were redirected to voicemail. This makes the call durations distribution to be significantly skewed towards smaller durations due to nontechnical failures, e.g., failure to answer. Finally, the authors showed that the distribution has a "semi-heavy" tail, with the variance being more than three times the mean, which is significantly higher than that of exponential distributions. Comparing to a log-normal distribution, though the tails agree better, they too diverge at large values, what asks for a more heavy-tailed distribution.

3 Calls Duration Distribution

3.1 Problem Definition

In this work, we analyze mobile phone records of 3.1 million customers during four months. In this period, more than 1 billion phone calls were registered and, for each phone call, we have information about the duration of the phone call, the date and time it occurred and encrypted values that represent the source and the destination of the call, that may be mobile or not. When not stated otherwise, the results shown in this work refer to the phone call records of the first month of our dataset. The results for the other 3 months are explicitly mentioned in Section 4.

The Call Duration Distribution (CDD) is the distribution of the call duration per user in a period of time, that in our case, is one month. In the literature, there is no consensus about what well known distribution should be used to model the CDD. There are researchers that claim that the PDD should be modeled by a log-normal distribution [7] and others that it should be modeled by the exponential distribution [19]. Thus, in this section, we tackle the following problem:

Problem 1. CDD FITTING. Given $d_1, d_2, ..., d_C$ durations of n_i phone calls made by a user i in a month, find the most suitable distribution for them and report its parameters.

As we mentioned before, there is no consensus about what well known distribution should be used to model the CDD, i.e., for some cases the log-normal fits well and for others, the exponential is the most appropriate distribution. Thus, finding another specific random distributions that could provide good fittings to a particular group of CDDs would just add another variable to Problem 1, without solving it. Therefore, we propose that the distribution that solves Problem 1 should necessarily obey to the following requirements:

- **R1**: Intuitively explain the intrinsic reasons behind the calls duration;
- **R2**: Provide good reliable fits for the great majority of the users.

In the following sections, we present a solution for Problem 1. In Section 3.2, we tackle Requirement $R1$ by presenting the TLAC model, that is a intuitive model to represent CDDs. Then, in Section 3.3, we tackle Requirement $R2$ by showing the goodness of fit of the TLAC model for our dataset.

3.2 TLAC Model

Given these constraints, we start solving Problem 1 by explaining the evolution of the calls duration by a survival analysis perspective. We consider that all the calls $c_1, c_2, ..., c_C$ made by a user in a month are individuals which are alive while they are active. When a phone call c_j starts, its initial lifetime $l_j = 1$ and, as time goes by, l_j progressively increments until the call is over. It is obvious that the final lifetime of every c_j would be its duration d_j.

In the survival analysis literature, an interesting survival model that can intuitively explain the lifetime, i.e. duration, of the phone calls is the *log-logistic* distribution. And besides its use in survival analysis [1,12,11], there are examples in the literature of the use of the log-logistic distribution to model the distribution of wealth [5], flood frequency analysis[14] and software reliability[6]. All of these examples present a modified version of the well known "rich gets richer" phenomenon. First, for a variable to be "rich", it has to face several risks of "dying" but, if it survives, it is more likely to get "richer" at every time. We propose that the same occurs for phone calls durations. After the initial risks of hanging up the call, e.g., wrong number calls, voice mail calls and short message calls such as "I am busy, talk to you later" or "I am here. Where are you?" type of calls, the call tends to get longer at every time. As an example, the lung cancer survival analysis case [1] parallels our environment if we substitute endurance to disease with propensity to talk: a patient/customer that has stayed alive/talking so far, will remain such, for more time, i.e., the longer is the duration of the call so far, the more the parties are enjoying the conversation and the more the call will survive.

Thus, to solve Problem 1, we propose the **Truncated Lazy Contractor (TLAC)** model, that is a truncated version of the log-logistic distribution, since it not contains the interval $[01)$. Firstly we show, in Figure 1-a, the Probability Density Function (PDF) of the TLAC , the log-normal and exponential distributions, in order to emphasize the main differences between these models. The parameters were chosen accordingly to a median call duration of 2 minutes for all distributions. The TLAC and log-normal distributions are very similar, but the TLAC is less concentrated in the median than the

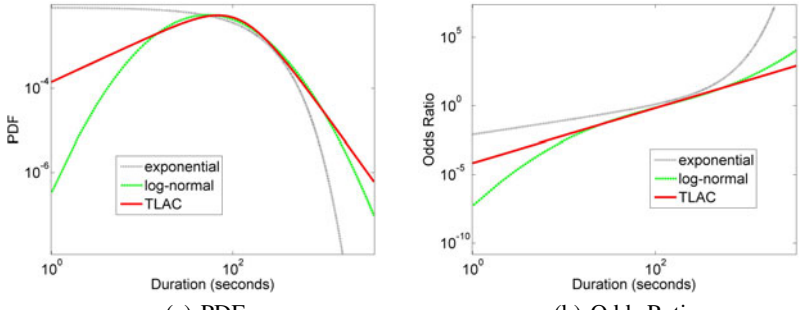
log-normal, i.e., it has power law increase ratios in its head and in its tail. We believe that this is another indication that the TLAC is suitable to model the users' CDD, since as it was verified by [20], CDDs have semi-heavy" tails. The basic formulas for the log-logistic distribution and, consequently, for the TLAC , are [11]:

$$PDF_{TLAC}(x) = \frac{\exp(z(1+\sigma)-\mu)}{(\sigma(1+e^z))^2}$$

$$CDF_{TLAC}(x) = \frac{1}{1+\exp(-\frac{(\ln(x)-\mu)}{\sigma})}$$

$$z = (\ln(x)-\mu)/\sigma$$

where μ is the location parameter and σ the shape parameter.



(a) PDF (b) Odds Ratio

Fig. 1. Comparison among the shapes of the log-normal, exponential and TLAC distributions

Moreover, in finite sparse data that spans for several orders of magnitude, that is the case of CDDs when they are measured in seconds, it is very difficult to visualize the PDFs, since the distribution is considerably noisy at its tail. One option is to smooth the data by reducing its magnitude by aggregating data into buckets, with the cost of lost of information. Another option is to move away from the PDF and analyze the cumulative distributions, i.e., cumulative density function (CDF) and complementary cumulative density function (CCDF) [2]. These distributions veil the sparsity of the data and also the possible irregularities that may occur for any particular reason. However, by using the CDF (CCDF) you end up losing the information in the tail (head) of the distribution. In order to escape from this drawbacks, we propose the use of the Odds Ratio (OR) function, that is a cumulative function where we can clearly see the distribution behavior either in the head and in the tail. This $OR(t)$ function is commonly used in the survival analysis and it measures the ratio between the number of individuals that have not survived by time t and the ones that survived. Its formula is given by:

$$OR(t) = \frac{CDF_{TLAC}(t)}{1-CDF_{TLAC}(t)} \tag{1}$$

Therefore, in Figure 2-b, we plot the OR function for the TLAC , the log-normal and exponential distributions. The OR function of the exponential distribution is a power

law until t reaches the median, and the it grows exponentially. On the other hand, the OR function of the log-normal grows slowly in the head and then fastly in the tail. Finally, the OR function for the TLAC is the most interesting one. When plotted in log-log scales, is a straight line, i.e., it is a power law. Thus, as shown in [1], the $OR(t)$ function can be summarized by the following linear regression model:

$$\ln(OR(t)) = \rho \ln(t) + \beta \tag{2}$$

$$OR(t) = e^{\beta} t^{\rho} \tag{3}$$

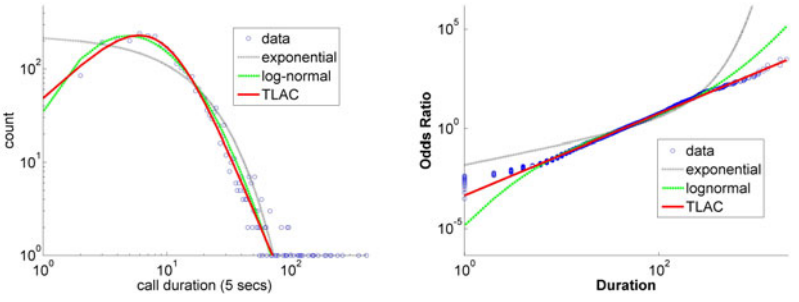
In our context, Equation 2 means that the ratio between the number of calls that will die by time t and the ones that will survive grows with a power of ρ. Moreover, given that the median $\hat{t}$ of the CDD is given when $OR(t) = 1$ and $OR(t) < 1$ when $t < \hat{t}$, the probability of a call to end grows with t when $t < \hat{t}$ and then decrease forever. We call this phenomenon the "lazy contractor" effect, which represents the time a lazy contractor takes to complete a job. If the job is easier and does require less effort than the ordinary regular job, he finishes it fastly. However, for jobs that are harder and that demand more work than the ordinary regular job, the contractor also gets more lazier and takes even more time to complete it, i.e., the longer a job is taking to be completed, the longer it will take. The ρ and the β are the parameters of the TLAC model, with $\rho = 1/\sigma$.

We conclude this section and, therefore, the first part of the solution to Problem 1, by explaining the intuition behind the parameters of the TLAC model. The parameter ρ is the *efficiency* coefficient, which measures how efficient is the contractor. The higher the ρ, the more efficient is the contractor and the faster he will complete the job. On the other hand, the location parameter β is the *weakness* coefficient, which gives the duration $\hat{t}$ of the typical regular job a contractor with a determined efficiency coefficient ρ can take without being lazy, where $\hat{t} = exp(-\beta/\rho)$. This means that the lower the β, the harder are the jobs that the contractor is used to handle.

3.3 Goodness of Fit

In this section, we tackle the second requirement of Problem 1 by showing the goodness of fit of our TLAC model. First, we show in Figure 2-a, the PDF of the CDD for a high talkative user, with 3091 calls, and with the values put in buckets of 5 seconds to ease the visualization. We also show the best fittings using Maximum Likelihood Estimation (MLE) for the exponential and the log-normal distributions and also for our proposed TLAC model. Visually, it is clear that the best fittings are the ones from the log-normal distribution and the TLAC distribution, with the exponential distribution not being able to explain either the head and the tail of the CDD.

However, by examining the OR plot in Figure 2-b, we clearly see the the TLAC model provide the best fitting for the real data. As verified for the exponential distribution in the PDF, in the OR case, the log-normal also could not explain either the head and the tail of the CDD. We also point out that we can see relevant differences between the TLAC model and the real data only for the first call durations, that happen because these regions represent only a very small fraction of the data. The results showed in Figure 2 once more validate our proposal that the TLAC is a good model for CDDs.



(a) PDF (b) Odds Ratio

Fig. 2. Comparison of models for the distribution of the phone calls duration of a high talkative user, with 3091 calls. TLAC in red, log-normal in green and exponential in black. Visually, for the PDF both the TLAC and the log-normal distribution provide good fits to the CDD but, for the OR, the TLAC clearly provide the best fit.

Given our initial analysis, we may state that the TLAC seems to be a good fit for the CDDs and also serve as an intuitively explanation for how the durations of the calls are generated. However, in order to conclude our answer for Problem 1, we must verify its generality power and also compare it to the log-normal and exponential generality power as well. Thus, we verify which one of the distributions can better fit the CDD of all the users of our dataset that have $n > 30$ phone calls. We calculated, for every user, the best fit according to the MLE for the TLAC , the log-normal and exponential distributions and we performed a Kolmogorov-Smirnov goodness of fit test [13], with 5% of significance level, to verify if the user's CDD is either one of these distributions. For now on, every time we mention that a distribution was correctly fitted, we are implying that we succesfully performed a Kolmogorov-Smirnov goodness of fit test.

In Figure 3, we show the percentage of CDDs that could be fitted by a log-normal, a TLAC and a exponential distribution. As we can see, the TLAC distribution can explain the highest fraction of the CDDs and the exponential distribution, the lowest. We observe that the TLAC distribution correctly fit almost 100% of the CDDs for users with $n < 1000$. From this point, the quality of the fittings starts to decay, but significantly later than the log-normal distribution. We emphasize that the great majority of users have $n < 1000$, what indicates that some of these talkative users' CDD are probably driven by non natural activities, such as spams, telemarketing or other strong comercial-driven intents. This result, allied to the fact that the TLAC distribution could model more than 96% of the users, make it reasonable to answer Problem 1 claiming that the TLAC distribution is the standard model for CDDs in our dataset.

Finally, we further explore Problem 1 by looking at the OR of the talkative users that were not correctly fitted by the TLAC model. In Figure 4, we show the OR for three of these users and, as we observe, even these customers have a visually good fitting to the TLAC model. These results corroborate even more with the generality power of TLAC . Despite of the fact that the irregularities of these customers' CDDs unable them to be correctly fitted by the TLAC model, it is clear that the TLAC can represent their CDDs significantly well.

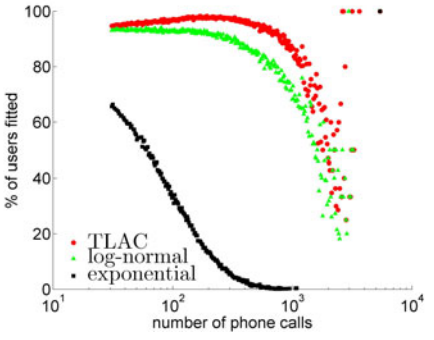
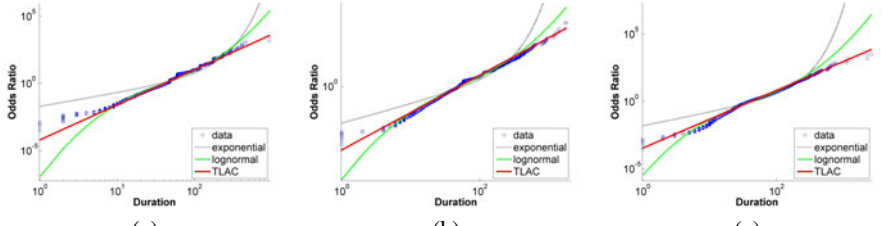


Fig. 3. Percentage of users' CDDs that were correctly fitted vs. the user's number of calls c. The TLAC distribution is the one that provided better fittings for the whole population of customers with $c > 30$. It correctly fitted more than 96% of the users, only significantly failing to fit users with $c > 10^3$, probably spammers, telemarketers or other non-normal behavior user.



(a) (b) (c)

Fig. 4. Odds ratio of 3 talkative customers that were not correctly fitted by the TLAC model

4 TLAC over Time

We know it is trivial to visualize the distribution of users with a determined summarized attribute, such as number of phone calls per month or aggregate calls duration. However, if we want to visualize the distribution and evolution of a temporal feature of the user such as his CDD, things start to get more complicated. Thus, in this section, we tackle the following problem:

Problem 2. EVOLUTION. Given the ρ_i and β_i parameters of N customers ($i = 1, 2..., N$), describe how they collectively evolve over time.

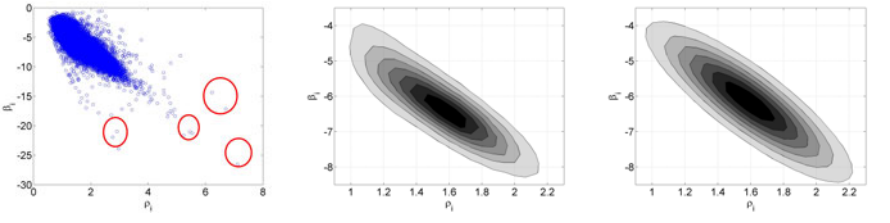
We propose two approaches to solve Problem 2. In Section 4.1 we describe the *MetaDist* solution and, in Section 4.2, we describe the *Focal Point* approach.

4.1 Group Behavior and Meta-fitting

Since we know that the great majority of users' CDD can be modeled by the TLAC model, in order to solve Problem 2, we need to figure out how each user i is distributed according to their parameters ρ_i and β_i of the TLAC model. If the meta-distribution

of the parameters ρ_i and β_i is well defined, then we can model the collective call behavior of the users and see its evolution over time. From now on, we will call the meta-distribution of the parameters ρ_i and β_i the *MetaDist* distribution.

In Figure 5-a, we show the scatter plot of the parameters ρ_i and β_i of the CDD of each user i for the first month of our dataset. We can not observe any latent pattern due to the overplotting but, however, we can spot outliers. Moreover, by plotting the ρ_i and β_i parameters using isocontours, as shown in Figure 5-b, we automatically smooth the visualization by desconsidering low populated regions. While darker colors mean a higher concentration of pairs ρ_i and β_i, white color mean that there are no users with CDDs with these values of ρ_i and β_i.



(a) Rough scatter plot (b) Isocontours of the real data (c) Bivariate Gaussian fitting.

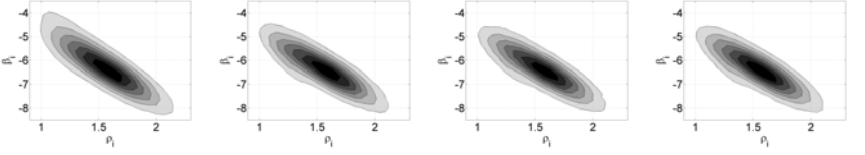
Fig. 5. Scatter plot of the parameters ρ_i and β_i of the CDD of each user i for the first month of our dataset. In (a) we can not see any particular pattern, but we can spot outliers. By plotting the isocontours (b), we can observe how well a bivariate Gaussian (c) fits the real distribution of the ρ_i and β_i of the CDDs ('meta-fitting').

Surprisingly, we observe that the isocontours of Figure 5-b are very similar to the ones of a bivariate Gaussian. In order to verify this, we extracted from the *MetaDist* distribution the means P and B of the parameters ρ_i and β_i, respectively, and also the covariance matrix Σ. We use these values to generate the isocontours of a bivariate Gaussian distribution and we plotted it in Figure 5-c. We observe that the isocontours of the generated bivariate Gaussian distribution are similar to the ones from the *MetaDist* distribution, which indicates that both distributions are also similar. Thus, we conjecture that a bivariate Gaussian distribution fits the real distribution of ρ and βs, making the *MetaDist* a good model to represent the population of users with a determined calls duration behavior.

Given that the *MetaDist* is a good model for the group behavior of the customers in our dataset, we can now visualize and measure how them evolve over time. In Figure 6 we show the evolution of the *MetaDist* over the four months of our dataset. The first observation we can make is that the bivariate Gaussian shape stands well during the whole analyzed period, what validates the robustness of the *MetaDist* . Moreover, a primarily view indicates that the meta-parameters also have not change significantly over the months. This can be confirmed by the first 5 rows of Table 1, which describes the value of the meta-parameters P, B and $\Sigma(\sigma^2_{\rho_i}, \sigma^2_{\beta_i}, cov(\rho_i, \beta_i))$ for the four analyzed months. This indicates that the phone company already reached a stable state before its customers concerning its prices, plans and services. In fact, the only noticeable

difference occurs between the first month and the others. We observe that the meta-parameters of the first month have a slightly higher variance than the others, what indicates that this is probably an atypical month for the residents of the country in which our phone records were collected. But in spite of that, in general, the meta-parameters do not change through time. Then, we can state the following observation:

Observation 1. *TYPICAL BEHAVIOR. The typical human behavior is to have a efficiency coefficient* $\rho \approx 1.59$ *and a weakness coefficient* $\beta \approx -6.25$*. Thus, the median duration for a typical mobile phone user is* 51 *seconds and the mode is* 20 *seconds.*



(a) Month 1 (b) Month 2. (c) Month 3 (d) Month 4.

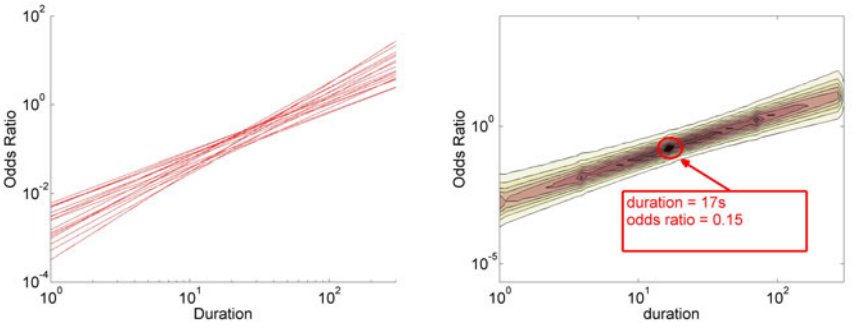
Fig. 6. Evolution of the *MetaDist* over the four months of our dataset. Note that the collective behavior of the customers is practically stable over time.

4.2 Focal Point

An interesting observation we can derive from the *MetaDist* showed in Figure 5 is that there exists a significant negative correlation between the parameters ρ_i and β_i. This negative correlation, more precisely of -0.86, lead us to the fact that the OR lines, i.e., the TLAC odds ratio plots of the customers of our dataset, when plotted together, should cross over a determined region. In order to verify this, we plotted in Figure 7-a the OR lines for some customers of our dataset. As we can observe, it appears that these lines are all crossing in the same region, when the duration is approximately 20 seconds and the odds ratio approximately 0.1. Then, in Figure 7-b, we plotted together the OR lines of $20,000$ randomly picked customers and derived from them the isocontours to show the most populated areas. As we can observe, there is a highly populated point when the duration is 17 seconds and the OR is 0.15. By analyzing the whole month dataset, we verified that more than 50% of the users have OR lines that cross this point. From now on, we call this point the *Focal Point*.

Formally, the *Focal Point* is a point on the OR plot with two coordinates: a coordinate $FP_{duration}$ in the duration axis and a coordinate FP_{OR} in the OR axis. When a set of customers have their OR plots crossing at a *Focal Point* with coordinates $(FP_{duration}, FP_{OR})$, it means that for all these customers the $\frac{FP_{OR}}{1+FP_{OR}}$th percentile of their CDD is on $FP_{duration}$ seconds. Thus, in the 2 bottom lines of Table 1, we describe the *Focal Point* coordinates for the four months of our analysis and, surprisingly, the *Focal Point* is stationary. Thus, we can make the following observation:

Observation 2. *UNIVERSAL PERCENTILE. The vast majority of mobile phone users has the same* 10*th percentile, that is on* 17 *seconds.*



(a) Direct plot. (b) Isocontours of the plot.

Fig. 7. The TLAC lines of several customers plotted together. We can observe that, given the negative correlation of the parameters ρ_i and β_i, that the lines tend to cross in one point (a). We plot the isocontours of the lines together and approximately 50% of the customers have TLAC lines that pass on the high density point (duration=17s, OR=0.15) (b).

Observation 2 suggests that one of the risks for a call to end acts in the same way for everyone. We conjecture that, given the 17 seconds durations, this is the risk of a call to reach the voice mail of the destination's mobile phone, i.e., the callee could not answer the call. The duration of this call involves listening to the voice mail record and leaving a message, what is coherent with the 17 seconds mark. It would be interesting to empirically verify the percentage of phone calls that reaches the voice mail and compare with the *Focal Point* result.

Table 1. Evolution of the meta-parameters (rows 1-5) and the *Focal Point* (rows 6-7) during the four months of our dataset

-	1st month	2nd month	3rd month	4th month
P	1.59	1.58	1.59	1.59
B	-6.16	-6.28	-6.32	-6.30
$\sigma^2_{\rho_i}$	0.095	0.086	0.084	0.083
$\sigma^2_{\beta_i}$	1.24	0.98	0.95	0.94
$cov(\rho_i, \beta_i)$	-0.30	-0.24	-0.24	-0.23
$FP_{duration}(s)$	17	17	17	17
FP_{OR}	0.15	0.12	0.11	0.11

5 Discussion

5.1 Practical Use

In the previous section, we showed the collective behavior of millions of mobile phone users is stationary over time. We described two approaches to do that, one based on the *MetaDist* and the other based on the *Focal Point* . The initial conclusions of both approaches are same. First, the collective behavior of our dataset is stable, i.e., it does

not change significantly over time. Second, we could see a slight difference between the first month and the others, indicating that this month is an atypical month in the year. We believe that these two approaches can succinctly and accurately aid the mobile phone companies to monitor the collective behavior of their customers over time.

Moreover, since we could successfully model more than 96% of the CDDs as a TLAC , a natural application of our models would be for anomaly detection and user classification. A mobile phone user that does not have a CDD that can be explained by the TLAC distribution is a potential user to be observed, since he has a distinct call behavior from the majority of the other users. To illustrate this, we show in Figure 8 a talkative node with a CDD that can not be modeled by a TLAC distribution. We observe that this node, indeed, has an atypical behavior, with his CDD having a noisy behavior from 10 to 100 seconds and also an impressive number of phone calls with duration around 1 hour (or 5×700 seconds). Moreover, another way to spot outliers is to check which users have a significant distance from the main cluster of the *MetaDist* . As we showed in Figure 5-c, this can be easily done even visually.

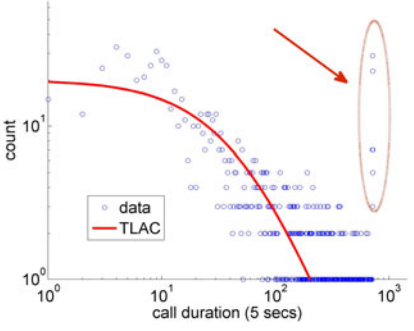


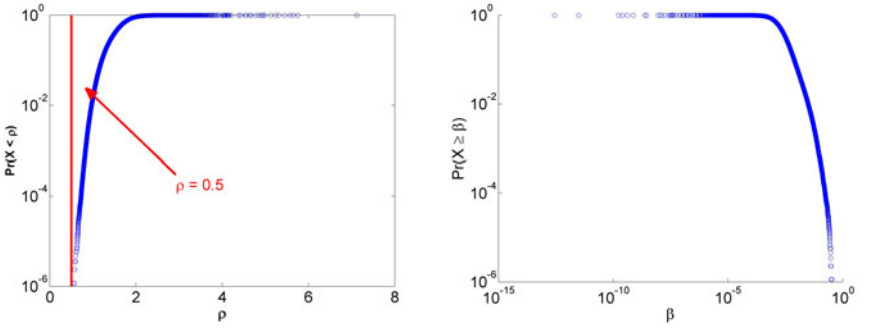
Fig. 8. Outlier whose CDD can not be modeled by the TLAC distribution

Another application that emerges naturally for our models is the summarization of data. By modeling the users' CDD into TLAC distributions, we are able to summarize, for each user i, hundreds or thousands of phone calls into just two values, the parameters ρ_i and β_i of the TLAC model. In our specific case, we could summarize over $0.1TB$ of phone calls data into less than $80MB$ of data. In this way, it is completely feasible to analyze several months, or even years of temporal phone calls data and verify how the behavior of the users is evolving through time. Also, all the proposed models in this work can be directly applied on the design of generators that produce synthetic data, allowing researchers that do not have access to real data to generate their own.

5.2 Generality of TLAC

As we mentioned earlier, one of the major strengths of the TLAC model is its generality power. We showed that even for distributions that oscillate between log-normal and log-logistic, or that have irregular spikes that unable them to be correctly fitted by TLAC , TLAC can represent them significantly well. Besides this, the simplicity of the TLAC

model allow us to directly understand its form when its parameters are changed and verify its boundaries. For instance, in the case of the CDD, e^{β} gives the odds ration when duration is 1 second. Thus, when $e^{\beta} > 1$, most of the calls have a lower duration than 1 second, which makes the CDD converges to a power law, i.e., the initial spike is truncated. Moreover, as $\alpha \to 0$, the odds ratio tends to be the constant e^{β}, what causes the variance to be infinity. By observing Figure 9 and concerning human calling behavior, we conjecture that β is upper bounded by 1 and ρ is lower bounded by 0.5. These values are coherent with the global intuition on human calling behavior.



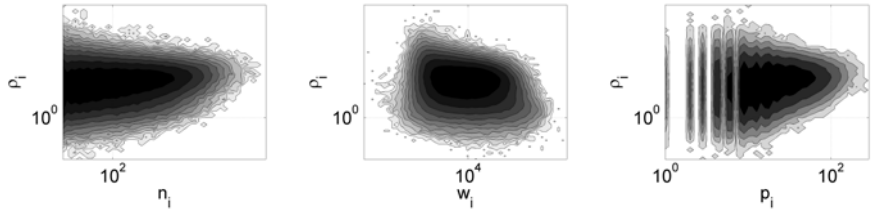
(a) CDF for ρ (b) CCDF for β

Fig. 9. Cumulative distributions for ρ and β. We can observe that ρ is lower bounded by 0.5 and β is upper bounded by 1. These values are coherent with the global intuition on human calling behavior.

5.3 Additional Correlations

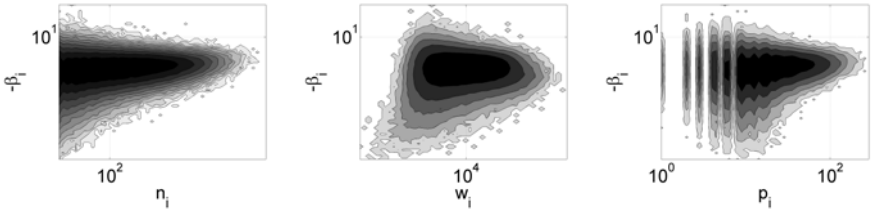
Given that the vast majority of users' CDDs can be represented by the TLAC model, it would be interesting if we could predict their parameters ρ_i and β_i based on one of their summarized attributes. One could imagine that a user that makes a large number of phone calls per month might have a distinct CDD than a user that makes only a few. Moreover, we could also think that a user that has many friends and talk to them by the phone regularly may also have a distinct CDD from a user that only talks to his family on the phone. In Figures 10 and 11, we show, respectively, the the isocontours of the behavior of the ρ_i and β_i parameters for users with different values of number of phone calls n_i, aggregate duration w_i and number of partners p_i, i.e., the distinct number of persons that the user called in a month. With the exception made for the ρ_i against w_i, we observe that the variance decreases as the value of the summarized attribute increases. This suggests that the CDD of high or long talkative users, as well as users with many partners, is easier to predict. Moreover, as we can observe in the figures and also in Table 2, there is no significant correlation between the TLAC parameters and the summarized attributes of the users. Thus, we make the following observation:

Observation 3. *INVARIANT BEHAVIOR. The ρ_i and β_i parameters of user i behave as invariant with respect to (a) number of phone calls n_i, (b) aggregate duration w_i and (c) number of partners p_i.*



(a) number of phone calls (b) aggregate duration (c) number of friends

Fig. 10. Isocontours of the users' CDD efficiency coefficient ρ and their summarized attributes



(a) number of phone calls (b) aggregate duration (c) number of friends

Fig. 11. Isocontours of the users' CDD efficiency coefficient β and their summarized attributes

Table 2. Correlations between summarized attributes and ρ and β

Attribute	Correlation with ρ	Correlation with β
number of phone calls	0.14	-0.18
aggregate duration	-0.21	0.01
number of partners	0.18	-0.18

Finally, since there is no significant correlation between the users' CDD parameters ρ_i and β_i with their summarized attributes, we emphasize that these parameters should be considered when characterizing user behavior in phone call networks. Moreover, besides characterizing individual customers, the TLAC model can also be directly applied to the relationship between users, analyzing how two persons call each other. One could use, for instance, the ρ parameter as the weight of the edges of the social network generated from phone call records.

6 Conclusions

In this paper, we explored the behavior of the calls' duration of the users of a large mobile company of a large city. We analyzed more than 3 million customers and 1 billion phone calls records. The main contributions of the paper are:

– The proposal of the TLAC distribution, which fits very well the vast majority of individual phone call durations, *much better* than log-normal and exponential;

- the introduction of *MetaDist* , which shows that the *collection* of TLAC parameters, and specifically the ρ and β ones, follow a striking bivariate Gaussian, with mean (P, B);
- Temporal evolution: the discovery that the *MetaDist* remains the same over time, with very small fluctuations;
- Usefulness of TLAC : it can spot anomalies (see Figure 8) and it can succinctly describe spot correlations (or lack thereof) between total phone call duration, number of calls, and number of distinct patterns, for a given user.

Moreover, we showed that TLAC has a very natural, intuitive explanation behind it (the more you waited so far, the even longer you will wait), and that it includes as special case the Pareto distribution.

Future work could focus on network effects, that is, if two people talk to each other, what is the relationship between their TLAC parameters? A second promising direction is to check whether TLAC also fits well other modes of human (or computer) communications, like length of SMS messages and length of postings on FaceBook "walls".

Acknowledgments. We thank the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for financial support. Research was also sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

1. Bennett, S.: Log-logistic regression models for survival data. Journal of the Royal Statistical Society. Series C (Applied Statistics) 32(2), 165–171 (1983)
2. Clauset, A., Shalizi, C.R., Newman, M.E.J.: Power-law distributions in empirical data. SIAM Review 51(4), 661 (2009), http://dx.doi.org/10.1137/070710111
3. Cortes, C., Pregibon, D., Volinsky, C.: Communities of interest. In: Hoffmann, F., Hand, D.J., Adams, N.M., Fisher, D.H., Guimarães, G. (eds.) IDA 2001. LNCS, vol. 2189, pp. 105–114. Springer, Heidelberg (2001)
4. Du, N., Faloutsos, C., Wang, B., Akoglu, L.: Large human communication networks: patterns and a utility-driven generator. In: KDD '09: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 269–278. ACM, New York (2009)
5. Fisk, P.R.: The graduation of income distributions. Econometrica 29(2), 171–185 (1961)
6. Gokhale, S.S., Trivedi, K.S.: Log-logistic software reliability growth model. In: HASE '98: The 3rd IEEE International Symposium on High-Assurance Systems Engineering, pp. 34–41. IEEE Computer Society Press, Washington (1998)
7. Guo, J., Liu, F., Zhu, Z.: Estimate the call duration distribution parameters in gsm system based on k-l divergence method. In: International Conference on Wireless Communications, Networking and Mobile Computing, WiCom 2007, pp. 2988–2991 (September 2007)

8. Hidalgo, C.A., Rodriguez-Sickert, C.: The dynamics of a mobile phone network. Physica A: Statistical Mechanics and its Applications 387(12), 3017–3024 (2008)
9. Hill, S., Nagle, A.: Social network signatures: A framework for re-identification in networked data and experimental results. In: CASON '09: Proceedings of the 2009 International Conference on Computational Aspects of Social Networks, pp. 88–97. IEEE Computer Society, Washington (2009)
10. Hill, S., Provost, F.J., Volinsky, C.: Learning and inference in massive social networks. In: Frasconi, P., Kersting, K., Tsuda, K. (eds.) MLG (2007)
11. Lawless, J.F.: Statistical Models and Methods for Lifetime Data (Wiley Series in Probability & Mathematical Statistics). John Wiley & Sons, Chichester (1982)
12. Mahmood, T.: Survival of newly founded businesses: A log-logistic model approach. JournalSmall Business Economics 14(3), 223–237 (2000)
13. Massey, F.J.: The kolmogorov-smirnov test for goodness of fit. Journal of the American Statistical Association 46(253), 68–78 (1951), http://dx.doi.org/10.2307/2280095
14. Ahmad, M.I., Sinclair, C.D., Werritty, A.: Log-logistic flood frequency analysis. Journal of Hydrology 98, 205–224 (1988)
15. Nanavati, A.A., Gurumurthy, S., Das, G., Chakraborty, D., Dasgupta, K., Mukherjea, S., Joshi, A.: On the structural properties of massive telecom call graphs: findings and implications. In: CIKM '06: Proceedings of the 15th ACM International Conference on Information and Knowledge Management, pp. 435–444. ACM, New York (2006)
16. Onnela, J.P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J., Barabási, A.L.: Structure and tie strengths in mobile communication networks. Proceedings of the National Academy of Sciences 104(18), 7332–7336 (2007)
17. Onnela, J.P., Saramaki, J., Hyvonen, J., Szabo, G., de Menezes, M.A., Kaski, K., Barabasi, A.L., Kertesz, J.: Analysis of a large-scale weighted network of one-to-one human communication. New Journal of Physics 9(6), 179 (2007)
18. Seshadri, M., Machiraju, S., Sridharan, A., Bolot, J., Faloutsos, C., Leskove, J.: Mobile call graphs: beyond power-law and lognormal distributions. In: KDD '08: Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 596–604. ACM, New York (2008)
19. Tejinder, S., Randhawa, S.H.: Network Management in Wired and Wireless Networks. Springer, New York (2003)
20. Willkomm, D., Machiraju, S., Bolot, J., Wolisz, A.: Primary users in cellular networks: A large-scale measurement study. In: 3rd IEEE Symposium on New Frontiers in Dynamic Spectrum Access Networks, DySPAN 2008, pp. 1–11 (2008)