ABSTRACT

How long are the phone calls of people using mobile phones? Here we answer this question by studying individual call duration distributions of users in large mobile networks. We analyzed a large, real network of 3.1 million users and 263.6 million phone calls from a private mobile phone company of a large city during one month. We discover surprising patterns for the call duration distribution of the users, and we propose the TLLOG distribution to fit them; TLLOG is the truncated version of so-called log-logistic distribution, a skewed, power-law-like distribution. We also propose the MetaDist to model the collective behavior of the users given their call duration distributions. The comparison between synthetic data and real data suggests that the MetaDist distribution accurately and succinctly describes the calls duration behavior of users in large mobile networks. Finally, as an additional contribution, we propose the Slice and Fit methodology to associate the call duration distribution of the users with their summarized features, such as aggregated calls duration, number of phone calls made and number of distinct partners.

1. INTRODUCTION

In the study of phone calls databases [17, 19, 14], a common technique to ease the analysis of the data is the summarization of the phone calls records into aggregated attributes [9], 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 [19], 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 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 Log-Logistic (TLLOG ) distribution to model how long are the durations of the phone calls of a single user. Thus, the TLLOG models the Calls Duration Distribution (CDD) of a user, and has only two parameters, the scale μ and the location σ. We show that the TLLOG distribution was the best alternative to model the CDD of the users of our dataset, mainly because it has a heavier tail than the log-normal distribution, that is the most commonly used distribution to model CDDs [6].

We also suggest the use of the TLLOG 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 distribution of the parameters μ and σ of each user’s CDD and, when visualized in logarithmic scales, its shape is surprisingly similar to a multivariate 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 TLLOG distribution to model the Calls’ Duration Distribution (CDD) of mobile users;
• The MetaDist to model group of customers and their CDDs;

As an additional contribution, we describe the Slice and Fit, a easy and simple de-aggregation methodology that enables the association of a summarized feature of a user to his CDD. We investigate the relationship between the number of calls, aggregate duration and the number of partners a user has with the μ and σ parameters of his CDD. By using the Slice and Fit methodology, we could achieve models that can accurately describe the duration of the calls of users with a determined summarized attribute. The are several applications that could benefit from the results presented in this paper. One could use the Slice and Fit methodology to summarize CDD data, and the MetaDist to generate synthetic datasets that contain information about the user’s CDD, and the TLLOG distribution to spot anomalies in a mobile network.

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 dataset and we show the TLLOG and the MetaDist distribution. The Slice and Fit de-aggregation methodology and the models of the users’ call behavior from their summarized attributes is shown in Section 4. In Section 5, we discuss 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 [9, 7]. In [15, 14], 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 [7], the authors verified that the persistence of an edge in highly correlated to its reciprocity and to the topological overlap and, in [3], 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 [8, 2], the authors proposed a framework and data structures for identifying fraudulent consuming on telecommunication networks based on their degree distribution and dynamics and, in [13], 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 [17], 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 [6], 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 [19], 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 $\psi$-semi-heavy-T 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. FITTING AND METAFITTING

In this work, we analyze mobile phone records of 3.1 million customers during one month. In this period, more than 263.6 million 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. In Figure 1, we plot the distribution of three summarized attributes of the users $i$ of our dataset. In Figure 1-a, we plot the distribution of the number of outgoing calls per user $c_i$ and in Figure 1-b, we plot the distribution of the aggregate duration of outgoing calls per user $d_i$. Finally, in Figure 1-c, we plot the distinct number of phones a user called per month $p_i$, that from now on will be referred as the number of partners. These shapes are similar to DPLN distributions reported in [17] but, in fact, there are still a lot of skewed distributions that we do not know what they are.

3.1 Fitting

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 [6] and others that it should be modeled by the exponential distribution [18]. Thus, in this section, we tackle the following problem:

**PROBLEM 1. CDD FITTING.** Let $C$ be a mobile phone company dataset with $n$ users $\{1, ..., n\}$ and $m$ phone calls records $\{r_1, ..., r_m\}$. Each phone call record $r$ is a triple $(i, j, w)$ in which $i$ is the caller, $j$ is callee and $w$ is the duration of the call. Find the best probability function $f_i$ that models the set of $c_i$ calls durations $C_i = \{w_i,1, ..., w_i,c_i\}$ for every user $i$.

In order to solve Problem 1, we propose the Truncated Log-logistic (TLLOG) distribution, that is a truncated version of the log-logistic distribution, since it not contains the interval $[01]$. A random variable is log-logistically distributed if the logarithm of the random variable is logistically distributed. The log-logistic distribution is very similar to the normal distribution, but it has heavier tails. In the literature, there are examples of the use of the log-logistic distribution in survival analysis [1, 10], distribution of wealth [4], flood frequency analysis[12] and software reliability[5]. The TLLOG distribution, as well as the log-logistic distribution, is a two-parameter distribution with parameters scale $\mu$ and shape $\sigma$ for values in the interval $(0, \infty)$. Its probability density function (PDF), its cumulative density function (CDF), its median and its range are given, respectively, by the following equations 1, 2, 3 and 4. We would like to emphasize that the parameter $\mu$ is related to the median of the distribution, not the mean.

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

$$CDF_{TLLOG}(x) = \frac{1}{1 + \exp(-\frac{(\ln(x) - \mu)}{\sigma})} \quad (2)$$

$$\hat{x}_{TLLOG} = \exp \mu \quad (3)$$

$$\hat{x}_{TLLOG} = \exp(\mu + \sigma \ln(\frac{1 - \sigma}{1 + \sigma})) \quad (4)$$

where

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

In Figure 2 we plot the CDD for two different users that have more than 1000 calls and we fit them using the TLLOG distribution by the Maximum Likelihood Estimation (MLE) method. Firstly, we can observe that their calls’ durations are very noisy and visually difficult to fit, confirming what was observed in [19]. Secondly, we observe that the fitted TLLOGs have different values of $\mu$ and $\sigma$. This was done to illustrate that as $\mu$ increases while $\sigma$ is kept the same, the PDF gets stretched out to the right, increasing the mode and decreasing its height, but maintaining its shape. On the other hand, as $\sigma$ decreases while $\mu$ is kept the same, the PDF gets pushed in towards the left, decreasing the mode and increasing its height.
We calculate their modes using Equation 4 and, for the user represented by circle dots, it is approximately $10 \times 5s = 50s$ seconds, while for the user represented by cross dots, it is approximately $12 \times 5s = 60s$ seconds. These values are, indeed, good estimators of the CDDs observed in Figure 2.

In order to conclude our answer for Problem 1, we verify which one of the distributions can better fit the CDD of all the users of our dataset that have $c > 30$. We calculated, for every user, the best fit according to the MLE for the log-normal, log-logistic and exponential distributions and we performed a Kolmogorov-Smirnov goodness of fit test [11], 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 successfully performed a Kolmogorov-Smirnov goodness of fit test.

In Figure 4-a, we show the percentage of CDDs that could be fitted by a log-normal, a TLLOG and a exponential distribution. As we can see, the TLLOG distribution can explain the highest fraction of the CDDs and the exponential distribution, the lowest. We observe that the TLLOG distribution correctly fit almost 100% of the CDDs for users with $c < 1000$. From this point, the quality of the fittings starts to decay, but significantly later than the log-normal distribution. Also, by observing again Figure 1-a, we see that the great majority of users have $c < 1000$, what indicates that some of these talkative users’ CDD are probably driven by non natural activities, such as spams, telemarketing or other strong commercial-driven intents. This result, allied to the fact that the TLLOG distribution could model more than 96% of the users, make it reasonable to answer Problem 1 claiming that the TLLOG distribution is the standard model for CDDs in our dataset. Moreover, we would like to point out that it should be at least be considered to represent the CDDs of other mobile calls data sets.

3.2 Grouped Call Duration Distribution

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 of a temporal feature of the user such as his PDD, things start to get more complicated. Thus, in this section, we tackle the following problem:

**Problem 2. METAFITTING.** Let $C$ be a mobile phone company dataset with $n$ users and each user $i$ has done, in period of
time $\Delta t$, $c_i$ phone calls with durations $C_i = \{w_{1i}, ..., w_{ci}\}$. Also, let the probability distribution function $f_i$ of $C_i$ determine the user $i$ calls duration behavior. Then, how can we model the population of users that have a determined behavior in their calls duration?

Since we know that the great majority of users’ CDD can be modeled by the TLLOG, we simply need to figure out how the users are distributed according to their parameters $\mu$ and $\sigma$ of the TLLOG distribution in order have the distribution of users with a determined call behavior. In Figure 5-a, we show, the isocontours of the distribution of the parameters $\mu$ and $\sigma$ of the TLLOG distributions that were correctly fitted. We call this distribution the MetaDist distribution. The colors go from red to dark blue, with red meaning a high concentration of pairs $\mu$ and $\sigma$ and with dark blue meaning that there are no users with CDDs with these values of $\mu$ and $\sigma$. We observe that the isocontours are very similar to the ones of a multivariate Gaussian, but since the plot is in logarithmic scales, we conjecture that these are isocontours of a multivariate log-normal distribution.

In order to verify this, we extracted from the $\mu$ and $\sigma$ pairs the mean and the variance of the logarithmic of the parameters $\mu$ and $\sigma$, and also their correlation, and we use these values to generate the isocontours of a multivariate Gaussian distribution in linear scales. We plotted this distribution in Figure 5-b. We observe that the isocontours of the generated bivariate Gaussian distribution are similar to the ones generated from the $\mu$ and $\sigma$ pairs solely, which indicates that both distributions are also similar. Thus, we conjecture that a bivariate (log)normal distribution fits the real distribution of $\mu$ and $\sigma$ of the CDDs (‘meta-fitting’) $\mu$ and $\sigma$ 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. Thus, in the next sections, we answer to the following problem:

**Problem 3. DE-AGGREGATION.** Given a set of $n$ users $\{1, ..., n\}$, each one with $k$ aggregated attributes $\{a_{1i}, ..., a_{ik}\}$, find functions $f(a_{ij})$ that associate each aggregate attribute $a_{ij}$, $1 \leq j \leq k$ of user $i$ with his CDD.

We solve this problem by modeling the behavior of the parameters $\mu$ and $\sigma$ of a user $i$ CDD accordingly to three of his summarized features: total number of phone calls $c_i$, aggregate call duration $d_i$, and number of partners $p_i$. From this, a $CDD_i$ of user $i$ may be directly calculated from one of these attributes $a_i$ in the following way:

$$CDD_i = \frac{\exp(z_a(1 + f_a^o(a_i)) - f_a^o(a_i))}{f_a^o(a_i)(1 + \exp(z_a))^2}$$

where

$$z_a = \frac{\ln(a_i) - f_a^o(a_i)}{f_a^o(a_i)}$$

**4. DE-AGGREGATION**

**4.1 Slice and Fit Methodology**

Given that the vast majority of users’ CDDs are TLLOG distributions, it would be interesting if we could predict their parameters
and $f^c(a_i)$ and $f^g(a_i)$ are, respectively, functions that model a summarized attribute $a_i$ of user $i$ into the $\mu_i$ and $\sigma_i$ of the user’s CDD. In summary, the Equation 5 is the PDF of the TLLOG distribution and $f^c_i(a_i)$ and $f^g_i(a_i)$ are, respectively, functions that associate $a_i$ with $\mu_i$ and $\sigma_i$.

Before explaining the models, we explain our methodology. We propose the Slice and Fit methodology to find the relationship between a summarized attribute of a user, such as his number of phone calls during a given time, with his CDD. We already know that the CDD of the users follow a TLLOG distribution with parameters $\mu$ and $\sigma$. If we can infer $\mu_i$ and $\sigma_i$ for a user $i$ from one of his summarized attributes $a_i$, we automatically have his CDD. Thus, we use the Slice and Fit methodology to find a model that describes the behavior of the users’ CDD parameters $\mu$ and $\sigma$ from one of their summarized attributes. The steps of Slice and Fit are described below and were naturally followed to extract the models showed in the next sections.

The Slice and Fit steps:

1. **Attribute Selection**: select a summarized attribute $a$ that may be related to the CDDs of the users;
2. **Slicing**: divide the range of values of $a$ from $\min(a)$ to $\max(a)$ into $n$ slices $s_1, \ldots, s_n$ and put user $i$ in slice $s_j$ if $a_i \geq s_j$ and $a_i < s_{j+1}$. Be sure that there is, in each slice, a statistically significant number of users.
3. **Sampling**: select a sample of random slices $s_j$ and correctly fit the distribution of the $\mu$s and $\sigma$s of every user $i \in s_j$ to a distribution $F$. Do not proceed if you cannot not find $F$;
4. **D-Fitting**: correctly fit all the slices $s$ to $F$ and store all the $K$ parameters $p_1, \ldots, p_K$ of $F$ for each slice $s$;
5. **C-Fitting**: find curves $f_{\rho_k}$ that describe the behavior of each parameter $\rho_k$ of $F$ according to each slice $s$;
6. **Linking**: link $a$ to $F$ and $F$ to $\mu$ and $\sigma$.

### 4.2 Number of Calls

The first summarized attribute we select to model is the number of phone calls a user made per month. In order to ease the comprehension, we will first apply the Slice and Fit methodology to the $\mu$ parameter and then to the $\sigma$ parameter. In Figure 6-a, we show the isocurves of the behavior of the $\mu_i$ parameter for users $i$ with a determined value of $c_i > 30$. The first thing we observe is what seems to be a uncorrelated regular behavior for the most frequent value $\mu$ according to the value of $c$. If we look at each vertical slice of Figure 6-a, that are defined by the number of calls $c$, we notice that $\mu$ values between $10^{0.55}$ and $10^{0.60}$ are the most frequent ones per each slice. This indicates that the median of the CDDs of the users in a month does not have a significant correlation with the number of phone calls they made. Moreover, the great majority of users’ CDD have a median between 30 and 50 seconds. We can also observe the distribution of the $\mu$ values per vertical slice changes significantly from the first $c$ values to the last ones. For instance, it is much more likely to have a user in $i$ with $c_i < 100$ calls per month with a $\mu_i < 10^{0.2}$ than a user $j$ with $c_j > 1000$ and $\mu_j < 10^{-2}$. If we accurately model the distribution of $\mu$ values given the $c$ values of the users, we can improve our understanding of how users with different attributes behave and, therefore, apply this knowledge in the design of several applications, such as outlier detection and user classification.

Thus, we proceed with the Slicing step of the Slice and Fit methodology dividing the distribution of Figure 6-a in linearly growing slices from $30 \leq c \leq 2000$, where the first slice has width 1 (30 $\leq c < 31$) and the last has width 100 (1900 $\leq c < 2000$). Next, we execute the Sampling step by examining some random slices and trying to fit them to a know distribution. In Figure 6-a we show the distribution of the $\mu$ values for the slice where all the users have $c = 100$. Coincidently, this distribution is correctly fitted by a log-logistic distribution with $\mu_c = 1.36$ and $\sigma_c = 0.06$. Note that we will use the subscripts $c$, $d$ and $f$ to differentiate the slices distribution parameters from the TLLOG distribution parameters.
Using the power law with slope $-0.1$ that fitted the values in Figure 7-b. With $g_c(c_i)$ and $g_h(c_i)$ defined, we now can estimate the $\mu_i$ value for a user $i$ with number of calls $c_i$.

$$h_c(c_i) = \exp(-0.008 \cdot \ln(c_i)^2 + 0.09 \ln(c_i) + 0.06)$$

(a) $\mu_c = \exp(-0.008 \cdot \ln(c_i)^2 + 0.09 \ln(c_i) + 0.06)$

(b) $\sigma_c = 0.1116c^{-0.1}$

Figure 7: User’s CDD $\mu$ follows a log-logistic distribution with parameters $\mu_c$ and $\sigma_c$. The evolution of these parameters given $c$ is explained by the curves showed in the figures.

With the $\mu$ of the CDD modeled, we now apply the Slice and Fit methodology to model the $\sigma$ of the CDD. Thus, we repeat the previous steps and we plot, in Figure 8-a, the $\sigma$ values for every user $i$ that has a value of $c_i > 30$. The result is very similar to the one in Figure 8-a but, in this case, the slices where correctly fitted by a Generalized Extreme Value distribution of the type I, i.e., that has its shape parameter $k = 0$ [16]. This distribution is also called Gumbel distribution and also has two parameters, location $\alpha_c$ and scale $\beta_c$. We can see its shape in Figure 8-b, when we computed the fitting for all $\sigma$ in the slice where users have $c = 100$.

As we did for the $\mu$ values, we need to find equations $u_c(c)$ and $v_c(c)$ that model the behavior of $\alpha_c$ and $\beta_c$ for values of all the slices that have $c > 30$. In a way that $\alpha_c = u_c(c)$ and $\beta_c = v_c(c)$. In Figure 9, we show $\alpha_c$ and $\beta_c$ for every value of $c > 30$ in order to extract $u_c(c)$ and $v_c(c)$. In Figure 9-a, we show that the relationship between $\alpha_c$ and $c$ is a power law with an extra parameter. The precise equation that defines $u_c(c)$ is described in the figure. Moreover, the equation that defines $v_c(c)$ is also a power law with the same extra parameter, and its complete description is showed in Figure 7-b. With $u_c(c)$ and $v_c(c)$ defined, we now can estimate the $\sigma_c$ value for a user $i$ with number of calls $c_i$.

It is interesting to point out that the $\mu_c$ values reaches his highest point at $g_h'(c_i) = 0 \rightarrow c > 250$ calls and then starts to decrease. On the other hand, the $\sigma_c$ values decreases at the same rate for all the values of $c$. This is coherent to the behavior of the distribution of the $\mu$ parameters showed in Figure 6-a, where the evolution of the slices width behave completely different in the intervals $30 \leq c < 250$ and $250 \leq c < 2000$. This, allied to the fact that the $\alpha_c$ and $\beta_c$ always decrease with $c$, indicates that there is a significant more variety of calls’ duration behaviors in the group of users with $c < 250$ than in the group of high talkative users.

We conclude the Slice and Fit methodology with the Linking step, by showing the complete models for the functions $f_c'(c_i)$ and $f_c''(c_i)$ of user $i$ with attribute $c_i$ number of calls per a period of time. They are given, respectively, by Equation 6, that is the PDF of the log-logistic distribution, and Equation 7, that is the PDF of the Gumbel distribution. These equations explain what is happening with the CDD of the users of our dataset given their number of calls. We suggest that telecom engineers can employ the Slice and Fit methodology to generate models just as like the ones described in Equations 6 and 7 when evaluating possible scenarios for a new service, or studying expansions for a current service. In general, the Slice and Fit methodology can play a fundamental role to different management functions in a telecom company.

$$f_c'(c_i) = \frac{\exp(z_c,\mu(1 + h_c(c_i)) - g_c(c_i))}{h_c(c_i)(1 + \exp(2z_c,\mu))^2}$$

(6)

$$f_c''(c_i) = \frac{1}{u_c(c_i)} \exp\left(\frac{c_i - v_c(c_i)}{u_c(c_i)} - \exp\left(\frac{c_i - v_c(c_i)}{u_c(c_i)}\right)\right)$$

(7)
we can observe that as the aggregate duration methodology in order to find associations between summarized features and the CDD of the users. However, we will not show the detailed mathematical models, since they are restrict to our dataset and should not be generalized. In Figure 10 we show the isocontours of the distribution of the $\mu_i$ and $\sigma_i$ for every user’s i CDD that have a determined aggregate duration $d_i > 600$ seconds. In Figure 10-a, we can observe that as the aggregate duration $d$ gets higher, the skewness of the distribution of the $\mu$ values for $d$ goes from negative to positive. Moreover, if we had shown Figure 10-a without the isocontours, one might think that there is a high positive correlation between the aggregate duration of a user and the median of his CDD. However, as we can observe, the correlation is indeed positive, but it is very small for the great majority of the users. We can only visualize a hight correlation if we only consider the users located in the border contours, the ones that are painted in tonalities of blue. However, they represent the minority of the customers and may contain several outliers. In Figure 10-b, we observe that the isocontours are more regular, what suggests that the distribution of customers with a determined value of $\sigma$ and $d$ has, in logarithmic scales, the shape of a bi-variate Gaussian.

4.3 Aggregate Duration

Both in this section and in the next one, we repeat the Slice and Fit methodology in order to find associations between summarized features and the CDD of the users. However, we will not show the detailed mathematical models, since they are restrict to our dataset and should not be generalized. In Figure 10 we show the isocontours of the distribution of the $\mu_i$ and $\sigma_i$ for every user’s i CDD that have a determined aggregate duration $d_i > 600$ seconds. In Figure 10-a, we can observe that as the aggregate duration $d$ gets higher, the skewness of the distribution of the $\mu$ values for $d$ goes from negative to positive. Moreover, if we had shown Figure 10-a without the isocontours, one might think that there is a high positive correlation between the aggregate duration of a user and the median of his CDD. However, as we can observe, the correlation is indeed positive, but it is very small for the great majority of the users. We can only visualize a hight correlation if we only consider the users located in the border contours, the ones that are painted in tonalities of blue. However, they represent the minority of the customers and may contain several outliers. In Figure 10-b, we observe that the isocontours are more regular, what suggests that the distribution of customers with a determined value of $\sigma$ and $d$ has, in logarithmic scales, the shape of a bi-variate Gaussian.

When analyzing every slice of the bi-variate distributions of Figure 10, we could correctly fit all of them to log-logistic distributions. The $\mu_i$ values were fitted to log-logistic distributions with parameters $\mu_{d,\mu}$ and $\sigma_{d,\mu}$ and the $\sigma_i$ values were fitted to log-logistic distributions with parameters $\mu_{d,\sigma}$ and $\sigma_{d,\sigma}$. We emphasize that the second subscript indicates the TLLOG parameter we are dealing with.

In Figure 11 we show the behavior of the parameters $\mu_{d,\mu}$, $\sigma_{d,\mu}$, $\mu_{d,\sigma}$ and $\sigma_{d,\sigma}$ when the aggregate duration $d$ is varied. It is interesting to observe in the behavior of the parameters $\mu_{d,\mu}$ (Figure 11-a) and $\sigma_{d,\mu}$ (Figure 11-b) that $\mu_{d,\mu}$ increases with $d$ while $\sigma$ stays the same for a wide range of $d$. This emphasizes and explains more the positive correlation verified between the aggregate duration and the median of the CDDs. Moreover, the $\mu_{d,\mu}$ curve has two inflection points, one near $d = 1500$ and other near $d = 60000$. These $d$ points are also verified in the curve that explains the $\mu_{d,\sigma}$ (Figure 11-c) parameter, suggesting that users with $d < 1500$, $1500 \leq d < 60000$ and $d \geq 60000$ may be different classes of users. We leave further explorations to future work.

4.4 Number of Partners

Finally, we use the Slice and Fit methodology to investigate, for every user $i$, the attribute number of partners $p_i$, that is the number of distinct phones user $i$ called in a period of time. In Figure 12, we show the behavior of the parameters $\mu$ and $\sigma$ for determined values of $p$. We observe that the bi-variate distributions are similar and, also, very noisy when $p$ is low. This indicates that is very hard to predict the CDDs based solely on the $p$ if $p$ is low. However, as $p$ increases, the distributions become more regular. Thus, by defining slices that contain only users with high values of $p$, we could correctly fit the distribution of the parameters $\mu_i$ and $\sigma_i$ as a log-logistic distribution for every user $i$ that has a number of partners $p_i > 44$.

Both $\mu$ and $\sigma$ distributions are, respectively, log-logistic distributions with parameters $(\mu_{p,\mu}, \sigma_{p,\mu})$ and $(\mu_{p,\sigma}, \sigma_{p,\sigma})$. In Figure 13, we show the behavior of the parameters $\mu_{d,\mu}$, $\sigma_{d,\mu}$, $\mu_{d,\sigma}$ and $\sigma_{d,\sigma}$ for values of $p > 44$. What is interesting about these parameters is that every one of them have a linear relationship with the number of partners $p$, all with negative slopes. This suggests that as the number of partners of a user increases, the less time he talks to them, since the $\mu$ of his CDD decreases and, consequently, the median.

5. METADIST AT WORK
Figure 11: Behavior of the parameters $\mu_d$, $\mu_d$, $\sigma_d$, $\sigma_d$ vs. $d$ when the aggregate duration $d$ is varied. We clearly see different trends when $d < 1500$, $1500 \leq d < 60000$ and $d \geq 60000$.

Figure 13: Behavior of the parameters $\mu_p$, $\mu_p$, $\sigma_p$, $\sigma_p$ vs. $p$ when the number of partners $p$ is varied. All of them show linear relationships with $p$ and have negative slopes.

Figure 12: Number of partners vs. log-logistic parameters. Both isocontours are noisy for low values of $p$, indicating that for these values $p$ is not correlated with the CDD.

Outlier detection.
Since we could successfully model more than 96% of the CDDs as a TLLOG, 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 TLLOG 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 14-a a talkative node with a CDD that can not be modeled by a TLLOG 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 $5 \times 700$ seconds).

Summarization.
Another application that emerges naturally for our models is the summarization of data. By modeling the users’ CDD into TLLOG distributions, we are able to summarize, for each user, hundreds or thousands of phone calls into just two values, the parameters $\mu$ and $\sigma$ of the TLLOG distribution. In our specific case, we could summarize 16GB of phone calls data into less than 50MB 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.

Finally, we emphasize that in the last decade, we saw a fast explosion in mobile services with a closer integration of voice, data, and multimedia applications. With the widespread penetration of Internet, the telecom world has become one of the fastest growing and changing markets. In this context, the Internet has become a key player for mobile phone users and has the potential to offer an unprecedented opportunity to versatile multimedia and information services and innovate the way they are created and provided. On the other hand, in a marketplace with a multitude of services, the addition of a new one is a daunting challenge and can be very difficult to evaluate both technically and economically for those who create,
deploy and manage the service. For instance, how can a telecom company harness a new technical opportunity? Part of this answer is to obtain as much information as possible about a potential service before deploying it to mobile users. In this case, telecom engineers can employ the Slice and Fit de-aggregation methodology in order to mathematically understand the calling behavior of their customers and, therefore, evaluate possible scenarios for a new service, study expansions for a new one or even terminate a service that is jeopardizing the company’s business.

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 260 million phone calls records. The main contributions of the paper can be summarized as follows:

- We propose the TLLOG distribution to model the Calls’ Duration Distribution (CDD) of mobile users. We showed that the TLLOG distribution is an interesting and, in our case, more appropriate alternative to model CDDs in comparison to the more often used log-normal and exponential distributions.
- We propose the MetaDist distribution to model group of customers and their CDDs. The MetaDist distribution models the distribution of the parameters of the users’ CDDs.

We also proposed the Slice and Fit de-aggregation methodology to associate summarized features of the users to the parameters of their CDDs. The Slice and Fit methodology, as long with the TLLOG and MetaDist distributions, may contribute significantly in the design of realistic generators of mobile phone synthetic data and also may significantly aid in the management of large telecom companies.

As future work, we intend to repeat this work for several consecutive months of data and also in different datasets in order to better understand the temporal evolution of the users’ CDDs. It would be interesting to know, for instance, if the TLLOG distribution is also a good model for other mobile networks and, case positive, what should be the shape of the MetaDist for this data. Moreover, it would be fascinating if we find similarities between the Slice and Fit models proposed in this work with other ones generated from other datasets.

7. REFERENCES