A Probabilistic Framework for Estimating Call Holding Time Distributions

Saliha Büyükoçak, Güneş Karabulut Kurt Member, IEEE, Okan Cengaver

Abstract—Call holding time (CHT) is a statistical indicator in a cellular network, directly affecting network performance and providing critical insight for the network service provider. CHT distribution estimation literature relies on the classical estimation theory that targets to determine parameters of a function. Hence related work can be considered as making use of parametric approaches. However, the required assumptions for these approaches may not be correct in order to obtain an accurate model. In this paper, we introduce a probabilistic framework for CHT distribution estimation, which makes use of Dirichlet process mixture of lognormal distributions. The purpose of this work is to provide a practical Bayesian inference framework to enable the extraction and identification of user behaviors, which are not available through traditional estimation techniques. The performance of the proposed framework is tested on a large dataset that is obtained from a mobile switching center of a cellular network service provider composed of calls from GSM and HSPA networks. Accuracy of the obtained CHT distributions are verified through several performance tests, showing that all distribution estimates have significance levels of 0.99.

Index Terms—Call holding time distribution, cellular radio, Dirichlet process mixture model, mixture of lognormal distributions.

I. INTRODUCTION

Identifying structural patterns of user behavior from network traffic is highly critical from any network service provider’s perspective. In today’s demand driven architectural challenges, this information can dramatically help traffic planning and network management. Statistical information about usage patterns and its implications in the network links can aid a service provider to characterize network resource usage and user behavior, to infer traffic demands, to detect traffic/usage anomalies and to design pricing policies. Call holding time (CHT) is one of these statistical indicators of network traffic, directly affecting performance metrics such as call blocking and dropping probabilities in cellular networks.

Exponential distribution, which played an important role in teletraffic analysis and performance evaluation of cellular networks dating back to times when wired communication was popular, was the preferred mathematical model for CHT distribution [1]. However, user behaviors have evolved in parallel with the advances experienced in communication systems. Accordingly, studies using analytical tools, simulations and real datasets have been conducted, consequently proving that the exponential distribution, which is frequently assumed in queuing systems, is insufficient. Majority of the related studies indicate that a better fit model for CHT distribution is a mixture of lognormal distributions. A classical estimation technique, maximum likelihood estimation (MLE), has been used in the literature to determine the unknown parameters of lognormal-k distribution, by assuming that the number of mixture components, k, is known (commonly 2 or 3). However, k and related parameters of these mixtures need to be accurately estimated in order to determine usage patterns in a network.

To achieve this objective, a Bayesian inference based CHT distribution estimation framework is proposed in our work. Although our model can be considered as non-parametric, we need to note that in the literature, these models are defined as probabilistic models where the number of parameters scale proportionally with the data. The terminology arises from the fact that the number of parameters to be estimated is unknown and may possibly be infinite. These models are helpful in order to eliminate any dependency on parametric assumptions, such as lognormal-3 models. In the proposed CHT distribution estimation framework, we first introduce a rule make pre-filtering that targets to decompose complete dataset into subsets of smaller sizes according to the context of the data. We then make use of Dirichlet process mixture (DPM) models of lognormal distributions. DPM models can be applied to any exchangeable dataset where the
number of mixture components in the data is unknown [2]. Bayesian nonparametric models have recently been used in a variety of machine learning problems including regression, classification, clustering, latent variable modeling, sequential modeling, image segmentation, source separation and grammar induction [3].

In this paper, we show that the number of lognormal mixture components in the CHT distribution and the unknown parameters that belong to these mixtures can be optimally estimated through DPM models by using Bayesian inference techniques. The usage of nonparametric methods to extract/analyze mixture components in the CHT distribution is proposed here for the first time in the literature. The proposed method has been implemented to a large real dataset obtained from a mobile switching center (MSC) of AVEA, a cellular network service provider based in Turkey. Both the Global System for Mobile Communications (GSM) and the High Speed Packet Access (HSPA) networks are connected to this MSC. The obtained dataset consists of identifiers indicating positions of users and observations that state conversation durations. It has been inferred that the lognormal-k probability density functions (pdfs) can optimally model this real dataset. The main contributions of our work which uses the DPM model for Bayesian inference can be listed as follows.

1) A probabilistic estimation framework is given in order to optimally determine CHT pdfs in a given cellular network. Existing methods impose strict assumptions on the pdf estimates, which may not always be correct.

2) The extraction/analysis of user behaviors is enabled by mapping CHTs to distinct mixtures as the outcome of the estimation framework. Such a mapping is not possible to obtain with classical estimation, however probabilistic estimation techniques generate such a mapping as one of the outputs of the analysis steps.

The fact that the methodology is practically applicable enables facilitation of network/traffic planning activities in addition to providing significant insights regarding user behaviors by mapping each conversation to a unique mixture component. The related work is given in Section II. By using a 20 day dataset, gathered from an MSC, CHT distribution estimates are obtained. The details regarding the dataset are given in Section III. The DPM model based estimation framework along with the prefiltering process is proposed in Section IV. The obtained pdf estimates and their accuracy levels (with a significance level of 0.99) have been detailed in Section V, where it is shown that CHT distributions can be modeled more accurately with more than three components depending on conversation durations, proving the fact that the number components may vary. Finally, conclusions are drawn in Section VI.

II. RELATED WORK

Conversation durations, an important statistical indicator to analyze user behaviors in a network, are modeled with exponential distribution at times when wired networks were used to provide voice transmission, without any mobility support. Since then, user behaviors have changed as traditional network infrastructures were replaced with cellular networks and mobile users emerged. However, exponential distribution was still in use to model CHTs. The studies conducted in this respect can be compiled according to their publication dates as follows. Guerin studied channel occupancy time distribution for cellular radio systems through two different approaches (analysis and simulations) and came to the conclusion that channel occupancy time can be modeled with an exponential distribution in both cases [4]. In the study performed by Jedrzychki and Leung by using actual data obtained from a cellular network, it has been shown that channel holding times can be better modeled with a lognormal distribution [5]. Fang et al. demonstrated that channel occupancy time can only be exponential in the event that cell residence time is exponentially distributed [6]. Authors further observed that the general channel occupancy time statistics can be optimally modeled with exponential distribution only when mobility is low.

By using measurement data obtained from cellular telephone systems, Barcelo and Jordan inferred that the most accurate distribution for CHT is a sum of lognormal distributions and that exponential distribution is not proper for channel holding times [7], [8]. Soong and Barria reached the conclusion that channel occupancy times can be represented by using a Coaxian model when cell residence time has Erlang or hyper-Erlang distribution [9]. Sharp et al. have discovered that CHT distribution can be best represented with a lognormal distribution and that the CHTs are uncorrelated [10].

Pattavina and Parini, by using measured data, have researched a mathematical model for call holding times and they have observed that lognormal-3 distribution is the most proper model for real data [1]. In the study carried out by Yavuz and Leung, by using real data for voice transmission in the cellular networks, it has been shown that both channel occupancy times and call holding times can be optimally modeled with a lognormal distribution independent from the mobility of users [11]. Chen et al. analyzed the call holding times through using experimental data obtained from
mobile voice over IP calls, which consequently led them to the fact that VoIP CHTs can be modeled with a mixture of two lognormal distributions [12]. Generalizing the exponential CHT assumption, Wang and Fan obtained closed form analytical expressions for CHT under unencumbered session time and cell dwell time distributions [13]. Considering heavy-tailed distributions for cell dwell times in cellular networks, Corral-Ruiz et al. derived the analytical expressions for CHT [14], [15]. Based on these works, it can easily be concluded that exponential distribution is not sufficient to model the CHTs. In our study a mixture of lognormal distributions (also a heavy-tailed distribution), has been selected as the best mathematical model for CHT distribution, in accordance with the results of the existing studies.

The abovementioned studies have investigated mathematical models for CHTs using analytical tools, simulations and real data. The maximum likelihood estimation (MLE) was used for the purpose of finding the parameters of the candidate distributions in the studies conducted by using experimental datasets [1], [7], [8], [11], [12]. However, MLE technique can only be used to determine the values of unknown parameters belonging to a definite probability distribution [16]. In other words, the number of mixture components and correspondingly the number of the unknown parameters are assumed to be known in MLE. However, this assumption may not always be correct. In order to avoid such an assumption, a probabilistic framework (where the number of mixtures and their parameters are unknown) has been proposed in this work, details of which are given in Section IV.

III. EMPIRICAL DATASET DESCRIPTION

The dataset which we used within the scope of our work regarding the conversation durations has been provided by AVEA, a cellular network service provider in Turkey. A large real dataset composed of over five million observations acquired from GSM and HSPA networks for 20 days, following 1st of December, 2011.\(^1\) The proposed probabilistic CHT distribution estimation framework can be used with other exchangeable large datasets as well.

The simplified conceptual network structure, where data regarding voice calls from GSM and HSPA networks are logged, is shown in Fig. 1. In this figure, cells are idealized using hexagons and defined as an area that is served by a single base station. Base stations are connected to a base station controller (BSC), regulating radio resource usage and performing other control functions. Multiple BSCs are connected to an MSC. An MSC functions as a network controller; supporting network services, internetworking towards other networks, performing mobility management and other high level requirements about resource management though network management entities. A single MSC is shown in the figure, although in nation-wide deployments several MSCs are frequently required. During the paging stage of a call set-up, users are paged within location areas that consist of geographical areas that are controlled by one or more BSCs. Hence location areas are larger geographical spaces when compared to cells. In the figure, two location areas are shown in order for conceptual demonstration.

Data logging point of conversation durations is the MSC, as shown in the figure. Total number of calls in this large dataset is 5,045,070, dating from December 1st to December 20th of 2011. The call setup time is not included in the CHT. The CHT resolution is based on seconds and the maximum conversation duration is 4,200 seconds. Dataset includes cell identifiers that the call has started and terminated, along with the location area that the call has started and terminated. There are 5,090 distinct cells and 60 location areas included in this dataset. Hence, mobility patterns (MPs) of users can be identified by using starting and terminating cells and also starting and terminating location areas.

Three types of mobility patterns (highly mobile, mobile and stationary) have been identified with the purpose of examining effects of mobility on user behavior within

\(^1\)Note that this dataset, logged from one of the MSCs in the network, may not be representative of the nation-wide network traffic and may change with time.
the scope of our work. Based on the fact that a location area is larger than a cell, voice service users regarding calls of which the starting and terminating location areas are different can be assessed as highly mobile users (MP-1). User behavior with the same starting and terminating location areas and with distinct starting and terminating cells can be considered as mobile (MP-2). Users with the same starting and terminating cells during the same conversation can be identified as stationary (MP-3). The pre-filtering strategy which we followed while dividing our data into subsets according to mobility pattern has been summarized in Table I. Mixture components, contained in the dataset in accordance with the patterns given in this table, can be determined by using the CHT distribution estimation framework, detailed in the next section.

**IV. PROBABILISTIC CHT DISTRIBUTION ESTIMATION FRAMEWORK**

Let call holding times be defined as a random variable, $Y$, and let $y = y_1, y_2, \ldots, y_N$ denote our dataset of realizations of $Y$, where $N$ represents the length of our dataset. Here, each $y_i$ represents the duration of the $i^{th}$ conversation. Our goal is to determine $f_Y(y)$, as a mixture of $k$ lognormal distributions, where $k$ is unknown.

According to the literature about CHT distributions, it has been concluded that lognormal-k is a better mathematical model for experimental CHTs. Hence, it can be said that each observation is lognormally distributed and corresponding pdf of mixture of lognormal distributions can be written as

$$f_Y(y) = \sum_{m=1}^{k} \frac{1}{y\sqrt{2\pi\sigma_m^2}} \exp\left(\frac{(\log(y) - \mu_m)^2}{2\sigma_m^2}\right),$$

where $\gamma_m$, $\mu_m$ and $\sigma_m^2$ are the weight, the mean and the variance of the $m^{th}$ mixture component for $m = 1, 2, \ldots, k$. Here, $\sum_{m=1}^{k} \gamma_m = 1$. In (1), $\log(\cdot)$ represents the natural logarithm. Noting that a lognormal pdf can be uniquely defined by mean and variance, the estimation framework in this section targets to find a total number of $3k$ parameters. First, the parameter $k$ needs to be determined. Then $k-1$ mixture weights (summing up to 1), $k$ mean and $k$ variance values need to be evaluated.

As aforementioned, solution approaches towards such estimation problems are referred to as nonparametric since the proper value of $k$ needs to be estimated as well. In the proposed CHT estimation framework, in addition to estimation of the values of $3k$ parameters, we also aim to determine the corresponding mixture that each observation is a part of. Such a mapping can not be done with classical estimation techniques such as MLE. This would aid us to cluster user behaviors based on their CHTs. In order to achieve these objectives, we propose a probabilistic estimation framework, as detailed in the following subsections. Furthermore we provide details about evaluation metrics that are also used for evaluating the tests results, in order to demonstrate the high accuracy levels of the obtained CHT distribution estimates.

**A. Rule Based Pre-Filtering**

In order to provide a generalized framework for CHT distribution estimation we first introduce a rule based pre-filtering step, that can divide the overall dataset into mutually exclusive, collectively exhaustive data subsets. Properties of these subsets can be determined based on the target application. Since all subsets will have well-defined properties (or physical implications), this pre-filtering step will help us decompose and distinguish CHT pdfs within each subset. As an example, in this paper we focus on mobility pattern based pre-filtering as given in Table I. Accordingly, CHT subsets represent three mobility patterns; highly mobile, mobile and stationary. Collectively they construct the complete dataset.

By making use of the rule based pre-filtering, the overall pdf $f_Y(y)$ can be constructed from the conditional pdfs that can represent inherent properties of data as

$$f_Y(y) = \sum_{l \in L} p_l f_Y(y | l),$$

where $f_Y(y | l)$ is the conditional pdf and $p_l$ is the corresponding frequency of the $l^{th}$ data subset. Here, $l \in L$ is the subset index, and $\sum_{l \in L} p_l = 1$. At the end of this step we obtain $L = |L|$ subsets. Each data subset contains $n_l$ observations, $y_{l,1:n_l} = (y_{l,1}, \ldots, y_{l,n_l})$. Note that $\sum_{l \in L} n_l = N$.

**B. Determination of Mixture Model Parameters**

The second step in our methodology is to decompose $f_Y(y | l)$ into lognormal mixture components. In order to solve this estimation problem, Dirichlet process mixture (DPM) models can be utilized. DPM models were introduced by [2] and are commonly used in Bayesian nonparametric and Bayesian semi-parametric methods. Since their instruction, these models have been one of the most widely used nonparametric probabilistic models [3], [17], [18].

Bayesian inference methods, making use of DPM models, assume that parameters specifying the observation model are random variables and observations are
exchangeable. However we need to state that, solely by assuming that observations are exchangeable, is not a modeling assumption, but merely a mathematical consequence based on data’s properties. Exchangeability reflects the assumption that variables are independent of their observation orders, although they may be dependent among themselves [3], [19]. Considering the physical implications, exchangeability implies independence of an observation (conversation duration) from its entry order in the dataset logs. Therefore, we can state that any CHT dataset is exchangeable. Note that exchangeability in the dataset logs. Therefore, we can state that any CHT dataset is exchangeable. However we need to state that, solely by assuming that observations are exchangeable, is not a modeling assumption, but merely a mathematical consequence based on data’s properties. Exchangeability reflects the assumption that variables are independent of their observation orders, although they may be dependent among themselves [3], [19]. Considering the physical implications, exchangeability implies independence of an observation (conversation duration) from its entry order in the dataset logs. Therefore, we can state that any CHT dataset is exchangeable. Note that exchangeability

<table>
<thead>
<tr>
<th>Pattern Identifier</th>
<th>Description</th>
<th>Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>MP-1</td>
<td>Highly mobile</td>
<td>Starting and terminating location areas are distinct.</td>
</tr>
<tr>
<td>MP-2</td>
<td>Mobile</td>
<td>Starting and terminating location areas are the same. Starting and terminating cells are distinct.</td>
</tr>
<tr>
<td>MP-3</td>
<td>Stationary</td>
<td>Starting and terminating cells are the same.</td>
</tr>
</tbody>
</table>

TABLE I
MOBILITY PATTERN DESCRIPTIONS

1) Dirichlet Process Mixture Model for Lognormal Mixtures: In order to determine the mixture components of a DPM model, Blackwell and MacQueen introduced a Polya-Urn formulation [21], which we use in the following model formulation. Considering the DPM model, each observation of $y_{l,1:n_l} = (y_{l,1}, \ldots, y_{l,n_l})$ belongs to a mixture component. For any given mixture, all observations belonging to that mixture are independent draws from the same distribution. Let an assignment $z_{1:n_l} = (z_{l,1}, \ldots, z_{l,n_l})$ be a vector of mixture labels, and $k_l$ be the number of distinct mixtures in the assignment $z_{1:n_l}$. Then $z_{l,i} \in \{1, \ldots, k_l\}$ for $i = 1, \ldots, n_l$, and each mixture contains at least one observation. Due to existence of $z_{l,i}$’s, our proposed CHT estimation framework enables us to determine the mixture of each observation, mapping observations to mixtures. Assuming that each observation is drawn from a member of a family of lognormal distributions, $\theta_{l,j} = (\mu_{l,j}, \sigma^2_{l,j})$, where $\mu_{l,j}$ is the mean and $\sigma^2_{l,j}$ is the variance, for $l = 1, \ldots, L$, $j = 1, \ldots, k_l$. Let $f_{y_{l,j}}(y_{l,i} | \theta_{l,j})$ denote the density of $y_{l,1:n_l}$ for a given value of $\theta_{l,1:k_l} = (\theta_{l,1}, \ldots, \theta_{l,k_l})$, where $\theta_{l,i} = \theta_{l,j}$ only when $i = j$.

The probability distribution for $y_{l,i}$ under a DPM model has a hierarchical form. DPM model assumes that the mixture parameters are iid. Conditional on the mixture label $z_{l,i}$, the observations are independent of each other [22]. Firstly, we have a Dirichlet based prior for $z_{l,i}$, $p(z_{l,i})$ of $j^{th}$ mixture. Then, conditional on there being $k_l$ mixtures under our assignment, we have $k_l$ parameters $\theta_{l,j}$, which are independently drawn from a known prior $p(\theta_{l,j})$. Finally, conditional on $z_{l,i}$ and $\theta_{l,j}$, observation $y_{l,i}$ is drawn independently from $f_Y(y_{l,i} | \theta_{l,z_{l,i}})$. Thus, we have the following joint density [2], [22]

$$f_Y(y_{l,i}, z_{l,i}, \theta_{l,i}) = p(z_{l,i}) \prod_{j=1}^{k_l} p(\theta_{l,j}) \prod_{i=1}^{n_l} f_Y(y_{l,i} | \theta_{l,z_{l,i}}).$$

(3)

Based on the joint density expression given in (3), we use the following proposition in order to identify and determine mixture components.

**Proposition 1.** Noting that $y_{l,i}$ is lognormal distributed, logarithm of each observation can be modeled as a univariate Gaussian distribution with unknown mean and variance. A Gaussian distribution can be obtained by taking the logarithm of each CHT observation with lognormal distribution. Hence it can be stated that for $X = \log(Y)$

$$X \sim f_X(x) = \sum_{i=1}^{n_l} \pi_{l,i} N(\mu_{l,j}, \sigma^2_{l,j})$$

$$i = 1, \ldots, n_l,$$ (4)

where $N(\cdot)$ represents Gaussian pdf and $\pi_{l,j}$’s are the mixing proportions (which are positive and sum to one), for $l \in \mathcal{L}$ and $j = 1, \ldots, k_l$.

**Proof of Proposition 1:** Assuming that $X$ is a Gaussian random variable, $Y$ has lognormal probability density function when $Y = \exp(X)$ [23]. Noting that all conversation durations are positive (that is $y_{l,i} > 0$, $\forall l$ and $\forall i$) the mapping $X = \log(Y)$ can be shown to be bijective. Hence, logarithms of observations of CHTs are composed of a mixture of Gaussian random variables.

Using Proposition 1 and assuming an inverse Gamma prior for $\sigma^2_{l,j}$ so that $s_{l,j} = 1/\sigma^2_{l,j}$, then it can be shown that [2]

$$p(s_{l,j}) = \frac{s_{l,j}^{-\alpha} s_{l,j}^{\alpha} \exp(-bs_{l,j})}{\Gamma(\alpha)},$$

(5)

where the shape parameter, $a$, and the scale parameter, $b$, are known. $\Gamma(\alpha)$ is the Gamma function. Conditional on $s_{l,j}$, $\mu_{l,j}$ has a normal prior distribution with mean
\( \eta \) and variance \( \tau / s_{l,j} \). Mixture parameters are iid with [22]

\[
\mu_{l,j} \sim \mathcal{N}(\eta, \tau / s_{l,j}),
\]

(6)

\[
\sigma_{l,j}^2 \sim \text{InvGamma}(a, b).
\]

(7)

In (7), \( \text{InvGamma}(\cdot) \) represents the inverse Gamma function.

After stating the relations between Gaussian and lognormal random variables and the corresponding prior distributions, next we explain how to determine, \( k_l \) the number of mixture components in the observation subset, \( y_{l,i} \).

2) Chinese Restaurant Process and Gibbs Sampler:

Bayesian nonparametric models target to determine the number of mixture components by assuming that it is infinite, while specifying the prior over all possible distinct labelings. The prior over these labelings is referred to as the Chinese restaurant process (CRP) [19]. The CRP provides a distribution over all possible labelings of the data [20], hence can be used to determine the number of mixtures in the lognormal mixture model.

The CRP is named based on the following metaphor. In a restaurant with an infinite number of tables, customers entering the restaurant can either sit at an empty table (referring to a new mixture assignment) or at a table where another customer was already sitting (referring to an existing mixture). The first customer sits at the first table. The second customer can either sit at the first table with probability \( \frac{\alpha}{\alpha+1} \) or the second table with probability \( \frac{1}{\alpha+1} \). Referred to as the concentration parameter, here \( \alpha \) is a positive constant determining the distributions of the mixture label assignments. The \( n^{th} \) customer entering the restaurant can sit at each of the occupied tables with probability proportional to the number of previous customers sitting at that particular table or can sit at the next unoccupied table with probability proportional to \( \alpha \). At any point in this CRP, the assignment of customers to tables defines a random labeling. As aforementioned, the prior for the vector of mixture label \( z_{l,i} \) (i.e. table assignments) is parameterized by \( \alpha \), and can be defined recursively by using

\[
p(z_{l,i+1} = j | z_{l,i}) = \begin{cases} 
n_{l,j}/(i + \alpha) & \text{for } j = 1, \ldots, k_{l,i} \\
\alpha/(i + \alpha) & \text{for } j = k_{l,i} + 1
\end{cases},
\]

(8)

where \( k_{l,i} \) is the number of mixtures in the assignment \( z_{l,i} \), and \( n_{l,j} \) is the number of observations that \( z_{l,i} \) assigns to mixture \( j \) [20]. Hence with CRP, we obtain a distribution over all possible distinct labelings amongst mixtures and their posterior probabilities.

Note that the mixture assignments under this distribution are exchangeable, implying that the order of customer entering the restaurant does not change the posterior probabilities of possible labelings. Since the CHTs are exchangeable, CRP can be used to determine all possible labelings and their posterior probabilities [22]. As CRP can be used to determine CHT distributions of \( y_{l,i} \) with the highest posterior probability, we can optimally model the experimental CHT data in the maximum a posteriori sense [24].

The prior distribution in (8), dealing with the problem of random sampling from a collection of conditional distributions, is often difficult to work out analytically, necessitating the development of Monte Carlo procedures [25]–[27]. To obtain the labeling with the maximum posterior probability, we use the Gibbs sampler. This is a technique for indirectly generating random variables from a marginal distribution without having to calculate the density explicitly [28], and it can be used to implement a practically realizable distribution estimation framework for CHTs.

The pseudo-code of the used Gibbs sampler is given in Fig. 2. Here, at each iteration conditional probabilities of each of the \( n_l \) observations (samples) are evaluated. Relative conditional probabilities for each CHT observation are calculated and posterior probabilities for all possible distinct labelings are stored though the iterations. The algorithm is terminated once the number of iterations reaches \( \text{MaxIteration} \). Hence, the mixture labeling with the highest posterior probability can be selected, making use of the optimality of the DPM in the maximum a posteriori sense.

1) Initialize \( z_{l,1}, \ldots, z_{l,n_l} \).
2) for \( \tau = 1 \rightarrow \text{MaxIteration} \) do
   Sample \( z_{l,1}^{(\tau+1)} \sim p(z_{l,1}|z_{l,2}^{(\tau)}, \ldots, z_{l,n_l}^{(\tau)}) \).
   Sample \( z_{l,2}^{(\tau+1)} \sim p(z_{l,2}|z_{l,1}^{(\tau)}, z_{l,3}^{(\tau)}, \ldots, z_{l,n_l}^{(\tau)}) \).
   \vdots
   Sample \( z_{l,1}^{(\tau+1)} \sim p(z_{l,1}|z_{l,1}^{(\tau+1)}, z_{l,3}^{(\tau)}, \ldots, z_{l,n_l}^{(\tau)}) \).
   \vdots
   Sample \( z_{l,n_l}^{(\tau+1)} \sim p(z_{l,n_l}|z_{l,1}^{(\tau+1)}, \ldots, z_{l,1}^{(\tau+1)}, z_{l,j-1}^{(\tau)}, \ldots, z_{l,n_l}^{(\tau)}) \).
3) end for
4) return \( z_l \) with maximum posterior probability.

Fig. 2. Pseudo-code of the Gibbs sampler.

3) Parameter Inference: Following CRP and Gibbs sampling, we can use the formulation of Gaussian mix-
ture model in order to determine distribution parameters of mixture components. Conditional on an assignment \( z_{l,i} \), the posterior distribution of \( s_{l,j} \) and \( \mu_{l,j} \) (the parameters for the \( j^{th} \) mixture) depend on \( n_{l,j} \), the number of observations in the \( j^{th} \) mixture, \( \bar{y}_{l,j} \) the mean of these observations, and \( \sigma^2_{l,j} \), the variance of these observations. Hence, we get that \( f_Y(s_{l,j} | n_{l,j}, \bar{y}_{l,j}, \sigma^2_{l,j}, s_{l,j}) \) is Gamma density with shape parameter \( a + n_{l,j}/2 \) and scale parameter \( b + n_{l,j}/2 \left( \sigma^2_{l,j} + \left( \bar{y}_{l,j} - \eta \right)^2 \right) \) [2]. As a consequence, estimation of unknown variance parameter as given in (9), for each mixture in (2). Conditional mean expression \( f_Y(\mu_{l,j} | n_{l,j}, \bar{y}_{l,j}, \sigma^2_{l,j}, s_{l,j}) \) can be shown to be a Gaussian density with mean and variance as given in (10). Hence (10) can be used to estimate the mean and variance of unknown mean parameter in each mixture. Frequencies of each mixture can be calculated by \( \pi_{l,j} = \frac{n_{l,j}}{n} \). At this final step, mixture parameters can be determined by using (9) and (10).

C. Evaluation Metrics

In order to verify distributional discrepancies of the pdf estimates, several different performance evaluation metrics can be used. These metrics can be classified as goodness-of-fit tests and error vector norm based techniques.

Goodness-of-fit tests are statistical hypothesis tests, which are used to determine whether a certain statistical distribution will be proper to characterize sample dataset [11]. There are various methods to evaluate goodness-of-fit, however, the most popular techniques are Chi-square and Kolmogorov-Smirnov (K-S) tests [10]. Binned data are used in Chi-Square test and consequently some information loss may occur [10]. However, K-S test prevents this time step (bin) dependency due to the fact that it uses the cumulative distribution function (cdf) instead of the pdf [7].

There is a linear relationship between the difficulty of use of goodness-of-fit tests and dataset size [10], [29], [30]. When the test is used on a large dataset, the result that the sample data does not fit the candidate distribution is generally observed (the null hypothesis is rejected). Therefore, some elements of the dataset are discarded while goodness-of-fit tests are implemented [10]. However, as will be shown in the following section, due to the strength of accurate modeling of CHT distribution by the DPM model based probabilistic estimation framework, these tests are implemented in our work without discarding any of the observations.

Error vector norm based techniques can be defined as methods which aim at measuring the norm of difference between the obtained and actual pdfs. The mean relative difference (MRD) and weighted mean relative difference (WMRD) are two metrics frequently used in the literature in order to compare the actual distribution of a real dataset and the expected distribution [31], [32]. MRD is generally used to determine the difference measurement between two pdfs, and is calculated by

\[
MRD = \frac{1}{T} \sum_t |c_t - \hat{c}_t| \times 0.5, \tag{11}
\]

where \( t \) represents the CHT, \( T \) is maximum CHT, \( c_t \) is the number of CHT observations of duration \( t \) and \( \hat{c}_t \) is our estimate of \( c_t \).

WMRD determines the distance measurement between two distributions independent from the size of dataset. The value of WMRD can be calculated as

\[
WMRD = \frac{\sum_t |c_t - \hat{c}_t|}{\sum_t (c_t - \hat{c}_t)} \times 0.5. \tag{12}
\]

Although there is not any standard about the distance between the measured dataset distribution and the candidate distribution, it will be suitable for many implementations to have WMRD value smaller than one [32]. In our work, WMRD values are lower than one for all estimates, as will be shown in the following section.

V. Empirical Results

In this section, the results we obtained by using the experimental dataset collected from GSM and HSPA networks are given. The relation between CHT and the number of samples is given for one hour call holding time in Fig. 3. It is clearly seen that there is a log-normal relationship between CHT and the number of
observations when Fig. 3 (a) is examined. In Fig. 3 (b), the maximum conversation duration is limited with 100 seconds in order to observe this relationship more closely. The figure shows us that the number of users increases between 10 to 60 seconds and the number of observations decreases as conversation duration increases. A histogram, which shows how the number of total samples are distributed according to days, is given in Fig. 4. From this figure, it can be observed that the business days are more intensive than the weekends and that Sundays (December 4, 11, and 18 of 2011) are less congested.

As emphasized in Figs. 3 and 4, our dataset is quite large, thus a database has been established in order to carry out the analysis. Rule based pre-filtering was carried out for the purpose of designating the CHT distribution of the relevant users regarding their mobility patterns (characteristics of which were provided in Table I). This pre-filtering process was executed on the complete dataset and on each day. MySQL program was used for the required inquiries. The daily frequencies of mobility patterns are given in Fig. 5, where we can see that the mobility patterns exhibit a similar behavior according to the number of calls. In other words, the comments about the complete dataset in Fig. 4 are also valid for data subsets. The mobility pattern frequencies ($p_l$) obtained as a consequence of pre-filtering, are provided in Table II.

It can be seen that the values/estimations regarding the mobility pattern group ratios in the daily data subsets and the values belonging to the complete dataset are similar. Based on this conclusion, it can be stated that the data subset belonging to any day may represent the complete dataset. Considering this, the data subset which belong to 1st of December, 2011 shall be examined as a characteristic representation of the complete dataset. The number of calls made within this day is 257, 217 and the maximum CHT is 4,200 seconds.

The hourly change of the number of calls are given Fig. 6, according to mobility patterns of 1st of December, 2011. From this figure we can see that the working hours are pretty crowded, and that conversation durations partly decrease in the evening and that the least amount of conversations take place around 5 AM.
TABLE II
FRACIONS OF MOBILITY PATTERNS OF DATASET. $p_1$, $p_2$ AND $p_3$ ARE FREQUENCIES OF MP-1, MP-2 AND MP-3, RESPECTIVELY

<table>
<thead>
<tr>
<th>Date</th>
<th>$p_1$ (%)</th>
<th>$p_2$ (%)</th>
<th>$p_3$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>from 1st of December, 2011 to 20th of December, 2011</td>
<td>7.67</td>
<td>22.22</td>
<td>70.09</td>
</tr>
<tr>
<td>1st of December, 2011</td>
<td>7.57</td>
<td>22.33</td>
<td>70.09</td>
</tr>
<tr>
<td>2nd of December, 2011</td>
<td>7.60</td>
<td>22.60</td>
<td>69.78</td>
</tr>
<tr>
<td>3rd of December, 2011</td>
<td>7.41</td>
<td>22.24</td>
<td>70.34</td>
</tr>
</tbody>
</table>

Fig. 6. Hourly variations of number of calls for 1st of December, 2011 observations.

The data subset of 1st of December, 2011 was examined by using the methodology proposed in Section IV with the purpose of determining CHT distributions. In the scope of this work, primarily the calls were classified according to their mobility patterns by pre-filtering, using the rules given in Table I.

In the following step, $k$ lognormal mixtures of call durations and their corresponding parameters are evaluated by using the DPM model. At the final step, the distributional discrepancies are quantified by using goodness-of-fit tests and error vector norm based techniques. In order to find a proper distribution to model the experimental dataset, the characteristics of the dataset to which K-S test can be applied to and the problems faced during the implementation are stated [12]. In [12]’s study, it is particularly stressed that it is not suitable to implement the K-S test to a distribution of which parameters are unknown. Consequently, the all-parameters-known version of the K-S goodness-of-fit test [7] has been used in our work. Besides, the boundary condition which is the greatest difficulty faced when implementing the Chi-square test, has been overcome due to and the optimality of the DPM model. We have observed that bin resolution is independent for the DPM model based pdf estimates. The test results are given in Table III. The significance level of the performed tests is set equal to 0.99. The fact that the pdfs have accurately modeled the dataset can easily be observed through evaluation metrics.

The data subset of MP-1 observations of 1st of December, 2011, calls with different starting/terminating location areas have been analyzed with the purpose of finding the number of mixtures and their parameters and the results obtained are given in Table IV. Detailed results about pdf parameters are sorted according to their increasing means, in all tables. The number of mixtures in group MP-1 has been found as $k = 3$ and the conclusion that the CHT distribution can be modeled through lognormal-3 has been drawn.

In Fig. 7, the pdf of the lognormal-3 distributed CHTs is shown, and the corresponding parameters are listed in Table IV. It can be seen from the figure that the experimental data can optimally be modeled with a lognormal-3. What we intend to emphasize here is that the number of lognormal mixtures in the dataset regarding highly mobile users, has been determined in a nonparametric fashion through the DPM model, without previously setting the number of components to 3. Furthermore, the assignment of $z_{1,1,19,475}$ will map
TABLE III
EVALUATION METRICS FOR PROBABILITY DENSITY FUNCTION ESTIMATES OF CHT FOR 1ST OF DECEMBER, 2011 OBSERVATIONS
SIGNIFICANCE LEVEL IS 99% FOR GOODNESS-OF-FIT TESTS

<table>
<thead>
<tr>
<th>Pattern Identifier</th>
<th>Chi-square Test</th>
<th>Kolmogorov-Smirnov Test</th>
<th>MRD</th>
<th>WMRD</th>
</tr>
</thead>
<tbody>
<tr>
<td>MP-1</td>
<td>Passed</td>
<td>Passed</td>
<td>1.9510 × 10⁻⁸</td>
<td>0.1773</td>
</tr>
<tr>
<td>MP-2</td>
<td>Passed</td>
<td>Passed</td>
<td>1.0481 × 10⁻⁸</td>
<td>0.0947</td>
</tr>
<tr>
<td>MP-3</td>
<td>Passed</td>
<td>Passed</td>
<td>1.0217 × 10⁻⁸</td>
<td>0.0697</td>
</tr>
</tbody>
</table>

TABLE IV
MIXTURE MODEL PARAMETERS OF PROBABILITY DENSITY FUNCTION OF CHT FOR MP-1 1ST OF DECEMBER, 2011 OBSERVATIONS

<table>
<thead>
<tr>
<th>Number of mixture components</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of calls</td>
<td>19,475</td>
</tr>
<tr>
<td>Mixture Id</td>
<td>j = 1</td>
</tr>
<tr>
<td>Mean Values (µ₁,j)</td>
<td>3.7514</td>
</tr>
<tr>
<td>Variances (σ²₁,j)</td>
<td>0.6306</td>
</tr>
<tr>
<td>Weights (π₁,j)</td>
<td>0.2998</td>
</tr>
<tr>
<td>Number of calls in each mixture (n₁,j)</td>
<td>5,838</td>
</tr>
</tbody>
</table>

Fig. 8. Probability density function of CHT for MP-2 1st of December, 2011 observations: (a) pdf of lognormal-5, (b) Detailed view of pdf.

The data subset for MP-3 of 1st of December, 2011 (the calls with the same starting/terminating cells) have also been analyzed and the obtained pdf’s parameters are given in Table VI. The number of mixtures in MP-3 has been determined as k = 8 and the conclusion that the CHT distribution for stationary users can be modeled through a lognormal-8 has been drawn for the corresponding data subset. The CHT distribution plot characterized by the usage of lognormal-8 is shown in Figure 9. It can be mentioned that this plot can optimally model the dataset, in line with the evidences provided in Table III. We can obtain information about which observation belongs to which mixture by using \( z_{2,1:57,440} \). Note that conversations of \( j = 1, 2 \) are extremely short (i.e. less than 1 seconds), corresponding to calling and hanging up almost instantaneously.

Existing techniques in the literature make use of parametric techniques where the number of mixture components are assumed to be known a priori. By making use of the proposed CHT estimation frameworks, we could calculate the number of lognormal mixtures in each mobility pattern. Once the number of the mixture components are now determined, classical estimation techniques that exists in the literature can be applied to obtain the CHT distributions. In order to demonstrate the performance of the DPM, we used an expectation maximization (EM) algorithm based distribution fitting since the closed form expression for the lognormal mix-

Copyright (c) 2013 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.
TABLE V
MIXTURE MODEL PARAMETERS OF PROBABILITY DENSITY FUNCTION OF CHT FOR MP-2 1ST OF DECEMBER, 2011 OBSERVATIONS

<table>
<thead>
<tr>
<th>Number of mixture components</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of calls</td>
<td>57,446</td>
</tr>
<tr>
<td>Mixture Id</td>
<td>j = 1</td>
</tr>
<tr>
<td>Mean Values ($\mu_{2,j}$)</td>
<td>0.4372</td>
</tr>
<tr>
<td>Variances ($\sigma_{2,j}^2$)</td>
<td>0.0444</td>
</tr>
<tr>
<td>Weights ($\pi_{2,j}$)</td>
<td>0.0063</td>
</tr>
<tr>
<td>Number of calls in each mixture ($n_{2,j}$)</td>
<td>364</td>
</tr>
</tbody>
</table>

TABLE VI
MIXTURE MODEL PARAMETERS OF PROBABILITY DENSITY FUNCTION OF CHT FOR MP-3 1ST OF DECEMBER, 2011 OBSERVATIONS

<table>
<thead>
<tr>
<th>Number of mixture components</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of calls</td>
<td>180,296</td>
</tr>
<tr>
<td>Mixture Id</td>
<td>j = 1</td>
</tr>
<tr>
<td>Mean Values ($\mu_{3,j}$)</td>
<td>0.3988</td>
</tr>
<tr>
<td>Variances ($\sigma_{3,j}^2$)</td>
<td>0.0496</td>
</tr>
<tr>
<td>Weights ($\pi_{3,j}$)</td>
<td>0.0034</td>
</tr>
<tr>
<td>Number of calls in each mixture ($n_{3,j}$)</td>
<td>618</td>
</tr>
</tbody>
</table>

Fig. 9. Probability density function of CHT for MP-3 1st of December, 2011 observations: (a) pdf of lognormal-8, (b) Detailed view of pdf.

In order to emphasize the effect of pre-filtering process, we also analyzed the complete dataset of 1st of December, 2011. In this study, the effect of mobility is investigated through post-filtering. As the outcome of our analysis, we have observed that a lognormal-6 distribution can accurately model this dataset. Corresponding mean, variance and weight values are shown in Table VIII. Later, the rules defined in Table I are applied to determine relative ratios of each mobility pattern, also given in Table VIII. According to these results one can see that relative ratios of CHT distributions follow a similar pattern, independent of the mobility patterns. In order to demonstrate this conclusion, we may observe extremely short calls ($j = 1$ mixtures) have the lowest percentage in all cases, with 17.91%, 18.43% and 23.39% for MP-1, MP-2 and MP-3, respectively. Medium length calls (such as $j = 3$ and $j = 4$ mixtures) have the highest ratio for all mobility patterns. Considering the longest call mixture ($j = 6$), we can see that 3.58%, 3.95% and 3.50% of calls are of MP-1, MP-2 and MP-3, respectively. Hence it can be concluded that mobility is not correlated with CHTs, as users are making use of provided voice services independent from their mobility.

VI. CONCLUSION

A probabilistic framework for CHT distribution estimation has been proposed in our work in order to find the optimum distribution and to enable network planners to analyze user behavior patterns. The empirical dataset used in the analysis has been obtained from an MSC of AVEA, to which GSM and HSPA networks are connected. Throughout our analysis, first different mobility patterns in the dataset have been obtained by using a rule based pre-filtering with the purpose of examining the effects of mobility on user behavior. Later, by making
TABLE VII
COMPARISON OF PERFORMANCE EVALUATION METRICS OF THE PROPOSED FRAMEWORK WITH THE EXISTING ESTIMATION TECHNIQUES

<table>
<thead>
<tr>
<th>Pattern Identifier</th>
<th>Lognormal mixture pdf (DPM)</th>
<th>Lognormal mixture pdf (EM)</th>
<th>Exponential pdf (MLE)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MRD</td>
<td>WMRD</td>
<td>MRD</td>
</tr>
<tr>
<td>---------------------</td>
<td>-----</td>
<td>------</td>
<td>-----</td>
</tr>
<tr>
<td>MP-1</td>
<td>1.9510 × 10−10</td>
<td>0.1773</td>
<td>2.2797 × 10−10</td>
</tr>
<tr>
<td>MP-2</td>
<td>1.0481 × 10−10</td>
<td>0.0947</td>
<td>1.7325 × 10−10</td>
</tr>
<tr>
<td>MP-3</td>
<td>1.9217 × 10−10</td>
<td>0.0697</td>
<td>4.9237 × 10−10</td>
</tr>
</tbody>
</table>

TABLE VIII
MIXTURE MODEL PARAMETERS OF PROBABILITY DENSITY FUNCTION OF CHT FOR 1ST OF DECEMBER, 2011 OBSERVATIONS AND CORRESPONDING FRACTIONS OF MOBILITY PATTERNS OBTAINED WITH POST-FILTERING

<table>
<thead>
<tr>
<th>Mixture Id</th>
<th>m = 1</th>
<th>m = 2</th>
<th>m = 3</th>
<th>m = 4</th>
<th>m = 5</th>
<th>m = 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Values (µ_m)</td>
<td>0.4403</td>
<td>2.5850</td>
<td>3.7611</td>
<td>5.3736</td>
<td>6.9343</td>
<td>8.0869</td>
</tr>
<tr>
<td>Variances (σ_m²)</td>
<td>0.0617</td>
<td>0.4690</td>
<td>0.6339</td>
<td>1.1844</td>
<td>0.3777</td>
<td>0.0643</td>
</tr>
<tr>
<td>Weights (γ_m)</td>
<td>0.0132</td>
<td>0.2182</td>
<td>0.3295</td>
<td>0.2987</td>
<td>0.1043</td>
<td>0.0361</td>
</tr>
</tbody>
</table>

ACKNOWLEDGMENT

The authors would like to thank Taylan Cemgil and Yener Ulker for their guidance and extremely helpful discussions about probabilistic methods.

REFERENCES

[13] X. Wang and P. Fan, “Channel holding time in wireless cellular communications with general distributed session time and dwell...


Okan Cengaver holds B.S. and M.Sc degrees in Electronic and Telecommunication Engineering from Istanbul Technical University, Turkey. He has around 18 years of experience in the Telecom and IT sector. In 1995, he started his career at Telsim, a GSM Operator in Turkey, as a Network Engineer. In 1996, Okan moved to research and development and worked for Netas/Nortel Networks in different locations including Istanbul, Los Angeles and Belfast as Senior R&D Test and Software Engineer respectively. In 2000, Okan continued his carrier in Telecommunication sector and joined Avea. In 2001, he has become Performance & Quality Manager and continued as IT Quality Manager, in 2004. In 2006, he was appointed as revenue assurance & fraud manager and then promoted to revenue assurance, fraud & security director in 2008. Okan served as the IT operation director responsible of whole IT operations, including billing/charging, CRM, sales, business intelligence and value added services from 2009 to 2013. Currently he is the Head of IT Department, Vodafone Turkey.

Salih Büyükkorak was born in Safranbolu, Turkey. She received the B.Sc degree with high honors in electronics and communication engineering from Istanbul Technical University, Istanbul, Turkey, in 2011. Currently, she is pursuing the M.Sc degree at the same university. She has been working at Wireless Communication Research Laboratory of Istanbul Technical University since 2010. Her research interests include MIMO channel modeling, traffic modeling, network analysis and Bayesian inference techniques, including Dirichlet process models.

Güneş Karabulut Kurt received the B.S. degree with high honors in electronics and electrical engineering from Bogazici University, Istanbul, Turkey, in 2000. She received her M.A.Sc. and Ph.D. degrees both in electrical engineering from the University of Ottawa, Ontario, Canada in 2002 and 2006, respectively. From 2000 to 2005 she was employed as a Research Assistant at CASP Group, University of Ottawa. She was with TenXc Wireless between 2005 and 2006, where she worked on location estimation and RFID systems. She was with Edgewater Computer Systems Inc. from 2006 to 2008 where she was working on high-bandwidth networking in aircraft and priority based signaling methodologies. From 2008 to 2010, she was with Turkcell R&D Applied Research and Technology, in Istanbul. Currently, she is an associate professor at Istanbul Technical University, where she has been with since 2010. Her research interests include sparse signal decomposition algorithms, multicarrier networks, traffic analysis and network planning/management.

Saliha Büyükcorak was born in Safranbolu, Turkey. She received the B.Sc degree with high honors in electronics and communication engineering from Istanbul Technical University, Istanbul, Turkey, in 2011. Currently, she is pursuing the M.Sc degree at the same university. She has been working at Wireless Communication Research Laboratory of Istanbul Technical University since 2010. Her research interests include MIMO channel modeling, traffic modeling, network analysis and Bayesian inference techniques, including Dirichlet process models.