Adequacy of Data for Characterizing Caller Behavior

Santi Phithakkitnukoon
Department of Computer Science & Engineering
University of North Texas
Denton, TX 76203 USA
santi@unt.edu

Ram Dantu
Department of Computer Science & Engineering
University of North Texas
Denton, TX 76203 USA
rdantu@unt.edu

ABSTRACT
The increase of advanced service offered by cellular networks draws lots of interest from researchers to study the networks and phone user behavior. From the phone user's point of view, we are interested in learning caller behavior. In this paper, we characterize caller behavior using probabilistic models based on caller's call arrival, inter-arrival, and talk time from the call logs. A probabilistic model is generally used to predict or estimate the future observation which is conditioned by a knowledge of the historical data. The question is how much historical data is adequate? We answer this question by presenting a technique to detect and compute the adequate amount of historical data to capture the caller behavior. In fact, this adequate amount of historical data is proved to be more relevant to the future caller behavior than considering the entire historical data and hence useful for constructing a predictive model for caller behavior. In addition, we show the improvement in the performance of a Call Predictor [16] when applying adequacy of data. For our analysis, we use the real-life call logs of 94 mobile users collected at MIT over nine months by the Reality Mining Project group. This paper extends our understanding of caller behavior. We believe that the results are useful in constructing a predictive model of a time series.

Categories and Subject Descriptors
H.4.3 [**Information System Application**]: Communications Applications.

General Terms
Algorithms, Experimentation, Human Factors.

Keywords
Caller, Single-peak, Multi-peak, Trace distance, Convergence time.

1. INTRODUCTION
Telecommunication device such as telephone has moved beyond being a mere technological object and has become an integral part of many people's social lives.

The 2nd SNA-KDD Workshop '08 (SNA-KDD'08), August 24, 2008 , Las Vegas, Nevada , USA . Copyright 2008 ACM 978-1-59593-848-0...$5.00.

This has had profound implications on both how people as individuals perceive communication as well as in the patterns of communication of humans as a society. Learning human behavior has always been the subject of interest in scientific fields (e.g. [19], [2], and [13]). There are also scientific reports in learning and characterizing user and network behavior (e.g. [1], [20], and [6]).

In communication systems, a user can be a "caller" who initiates communication or a "callee" who receives request for a communication from caller. As a callee in a phone network, a user generally has received calls from several callers. We are interested in learning caller behavior. A knowledge of caller behavior can lead to a predictive model which forecasts or predicts the future behavior of the caller such as calling time and hence useful for scheduling and planning (e.g., it can be used to avoid unwanted calls and schedule time for wanted calls). It can also be useful for the Public Safety Answering Point (PSAP) for predicting 9-1-1 (emergency) calls. It can also be beneficial to voice spam detection and prevention, as well as call centers for resource utilization.

Predictive models derived from communication logs have been studied extensively (e.g. [22], [17], and [9]). Recently there has been growing interests in the field of mobile social networks analysis to study human behavior by combing the computer technology and social networks (e.g. [6], [5], and [7]), but due to the unavailability of data, there have been far fewer studies. The Reality Mining Project at Massachusetts Institute of Technology [12] has made publicly available large datasets which we use for our analysis in this paper.

Motivation

In [16], authors proposed a Call Predictor which made the next-24-hour incoming call prediction based on caller behavior and reciprocity which were extracted from call history. This raises a question of how much call history is actually needed. Does it mean the more historical data, the better performance of the predictor? To answer this question, we find it interesting to study caller behavior and the adequacy of caller's past history.

Main Contribution

The main contribution of this paper is to infer the adequacy of historical call data to capture the behavior of the caller in order to construct a predictive model for future behavior observation.

The rest of this paper is structured as follows: Section 2 describes and statistical analyzes the real-life datasets. Section 3 proposes the concept of adequacy of historical data and its computation. Section 4 carries out the validation of hypothesis. The paper is

concluded in section 5 with a summary and an outlook on future work.

2. REAL-LIFE DATASET AND ANALYSIS

In our daily life, we receive phone calls from family members, friends, colleagues, supervisors, neighbors, and strangers. We believe that every caller exhibits a unique calling pattern which characterizes the caller behavior.

To study the caller behavior, we use the real-life datasets of 94 individual call logs over nine months of the mobile phone users which were collected at Massachusetts Institute of Technology (MIT) by the Reality Mining Project [12]. These 94 individuals are faculties, staffs, and students. The datasets include people with different types of calling patterns and call distributions.

Each call record in the datasets has the 5-tuple information which includes:

- Date (date of call)
- Start time (start time of call)
- Type (type of call i.e., "Incoming" or "Outgoing")
- Call ID (caller/calee identifier)
- Talk time (duration of call).

We use the call logs to derive the traffic profiles for each caller by inferring the Arrival time (time of receiving call from the caller), Inter-arrival time (elapsed time between adjacent incoming calls from the caller), and Talk time (duration of call from the caller).

2.1 Arrival Time

Based on our real-life datasets of 94 mobile phone users with more than 2,000 combined callers, we can divide callers into two categories namely Single-peak callers and Multi-peak callers based on their arrival time.

2.1.1 Single-peak Callers

The single-peak callers are callers who tend to make more calls at around one particular time of the day and less and less number of calls as time of the call deviates from that time (favorite time). Thus, we make a hypothesis that call arrival time has a normal distribution $N(\mu,\sigma^2)$ where μ is the mean and σ^2 is the variance of call arrival time which can be calculated by Eq. (1) and Eq. (2) respectively.

$$\mu = \frac{1}{N}\sum_{n=1}^{N} w(n)\,, \tag{1}$$

$$\sigma^2 = \frac{1}{N}\sum_{n=1}^{N}\left(w(n)-\mu\right)^2\,. \tag{2}$$

The arrival time is now treated as a random variable X that consists of number of small random variables $\{x(1), x(2), x(3), \ldots, x(N)\}$ where N is the total number of calls and $x(n)$ is the n^{th} call arrival time, is normal random variable which has probability density function (pdf) given by Eq. (3).

$$f_x(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}\,. \tag{3}$$

Hence the probability of receiving a call from caller k at time x is given by Eq. (4), where μ_k and $\sigma_k{}^2$ are the corresponding mean and variance of call arrival time of caller k.

$$\Pr\{X_k = x\} = \frac{1}{\sqrt{2\pi\sigma_k{}^2}} e^{-(x-\mu_k)^2/2\sigma_k{}^2}\,, \tag{4}$$

To check our hypothesis, we randomly select 100 callers from our dataset and perform the chi-square goodness-of-fit test (or χ^2-test) [11] (for testing the validity of the assumed distribution for a random phenomenon). We find that 30 callers have normal distribution at significant level $\alpha = 0.1$. Therefore, these 30 callers are considered as single-peak callers and the other 70 callers who do not pass the χ^2-test then belong to another group of callers which will be described in the next section.

As an example, in Fig. 1 the histogram of the call arrival time on time-of-the-day scales of a single-peak caller and fitted normal distribution are illustrated where we shift our window of observation to begin at 5AM and end at 4:59AM such that the entire calling pattern is captured in the middle. In fact, we find that it is a proper window of observation for the majority of the callers in our datasets.

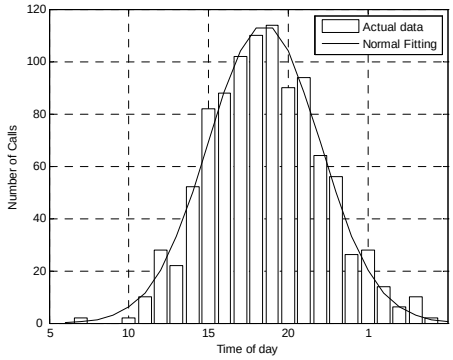

Figure 1: An example of single-peak caller whose call arrival time is fitted with normal distribution.

2.1.2 Multi-peak Callers

There is another group of callers whose calling behaviors based on arrival time are more random in the sense that they tend to have more than one favorite time of calling which result in more than one peak in their arrival time histograms.

The normal distribution is obviously not suitable for this type of callers. In fact, none of the parametric probability models fit to their structures. Therefore, probability density model must be determined from the data by using nonparametric density estimation. The most popular method for density estimation is the kernel density estimation (also known as the Parzen window estimator [14]) which is given by Eq. (5).

$$\hat{p}(x) = \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x-x_i}{h}\right)\,. \tag{5}$$

$K(u)$ is kernel function and h is the bandwidth or smoothing parameter. The most widely used kernel is the Gaussian of zero mean and unit variance which is defined by Eq. (6).

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\,. \tag{6}$$

The choice of the bandwidth h is crucial. Several optimal bandwidth selection techniques have been proposed ([10], [23]). In this paper, we use AMISE optimal bandwidth selection using the Sheather Jones Solve-the-equation plug-in method [21].

Likewise, the probability of receiving a call from caller k at time x can be computed similarly to Eq. (4) but using probability density function defined in Eq. (5).

As an example, the observed frequency of calls over nine months on time-of-day scales and fitted kernel density estimation are illustrated in Fig. 2.

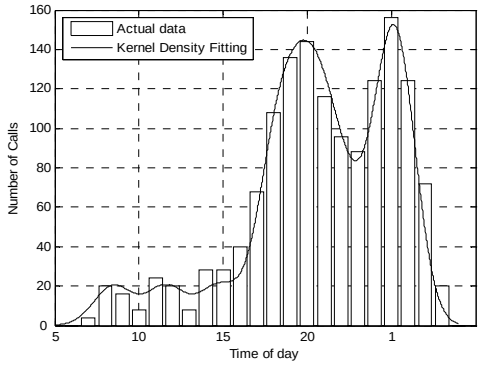

Figure 2: An example of multi-peak caller whose call arrival time is fitted with kernel density estimation.

2.2 Inter-arrival Time

Caller behavior can also be characterized by the inter-arrival time which is the time interval between adjacent incoming calls as it is monitored from the callee's point of view. Based on our dataset, by observing histograms of the inter-arrival time of all callers we find that they exhibit similar patterns in which the call frequency distribution is peaked at one particular point and exponentially decreases as inter-arrival time increases. Thus, we make a hypothesis that caller's inter-arrival time has an exponential distribution $exp(\gamma)$ where parameter γ is the rate at which calls are received. The parameter γ can be calculated by Eq. (7) and $E[Z]$ is the expected value of a random variable Z.

$$\gamma = \frac{1}{E[Z]}, \tag{7}$$

where inter-arrival time is a random variable Z which consists of small random variables $\{z(1), z(2), z(3), \ldots, z(N)\}$, where N is the total number of calls and $z(n)$ is the inter-arrival time of the nth call, i.e. interval of time from $(n\text{-}1)^{\text{th}}$ to n^{th} call. The pdf is given by Eq. (8).

$$f_Z(z) = \gamma e^{-\gamma z}, \tag{8}$$

Hence the probability of inter-arrival time from caller k is z time unit can be calculated by Eq. (9) where γ_k is the corresponding parameter of inter-arrival time of caller k.

$$\Pr\{Z_k = z\} = \gamma_k e^{-\gamma_k z}. \tag{9}$$

The chi-square goodness-of-fit test is also performed here to validate our hypothesis of assuming exponential distribution for caller's inter-arrival time. The tests are done using a significant level $\alpha = 0.1$ at which all callers pass the test and therefore confirm our hypothesis.

As an example, the histogram of inter-arrival time over nine months and fitted exponential distribution are illustrated in Fig. 3.

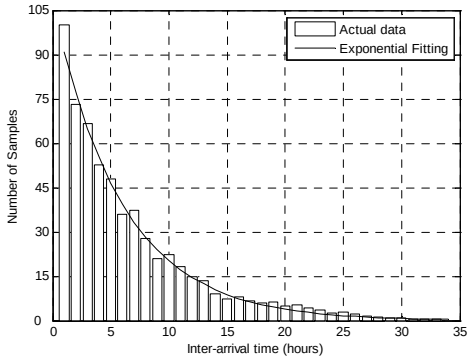

Figure 3: An example of caller's inter-arrival time is fitted with exponential distribution.

2.3 Talk Time

Talk time is the amount of time spent by the caller and callee during the call. From the callee's perspective, caller behavior can also be characterized by the talk time. Based on our observation of the histograms of the talk time of each caller, talk time exhibits an exponential-like pattern. Similar to the inter-arrival time pattern, the exponential distribution $exp(\lambda)$ is initially assumed for the talk time as our hypothesis where parameter λ can be calculated by Eq. (10) and $E[Y]$ is the expected value of a random variable Y.

$$\lambda = \frac{1}{E[Y]}. \tag{10}$$

Random variable Y represents the talk time that consists of small random variables $\{y(1), y(2), y(3), \ldots, y(N)\}$, where N is the total number of calls and $y(n)$ is the talk time of the n^{th} call. The pdf is given by Eq. (11).

$$f_Y(y) = \lambda e^{-\gamma y}. \tag{11}$$

Hence the probability of talk time with caller k is y time unit can be calculated by Eq. (12) where λ_k is the corresponding parameter of talk time of caller k.

$$\Pr\{Y_k = y\} = \lambda_k e^{-\lambda_k y}. \tag{12}$$

Similar to our previous cases, the chi-square goodness-of-fit test is also performed using a significant level $\alpha = 0.1$ at which all trials pass the test and therefore confirm our observation and hypothesis for talk time.

An example of a histogram of talk time over nine months of a sample caller who is randomly selected from our datasets and fitted exponential distribution is illustrated in Fig. 4.

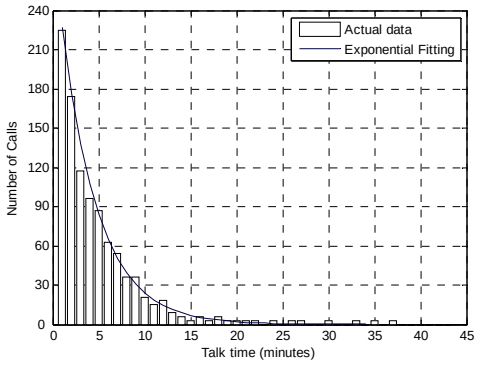

Figure 4: An example of caller's talk time is fitted with exponential distribution.

3. ADEQUACY OF HISTORICAL DATA

The caller behavior based on arrival time, inter-arrival time, and talk time have been characterized in forms of probability models in the previous section. Generally, a probability model is used to predict or estimate the future observation which is conditioned by a knowledge of the historical data. *The question is how much historical data is adequate?* This section attempts to answer this question.

In our case, the historical data is a collection of call logs which is a time series (a collection of observations made sequentially through time [3]). Unfortunately, the call logs are not deterministic (or can be predicted exactly) but stochastic in that future is only partly determined by past values, so that the exact predictions of future values are not quite possible and hence have a probability distribution.

The previous section shows that a single-peak caller can be characterized by a normal distribution model $N(\mu,\sigma^2)$ which is characterized be the mean μ and variance σ^2. In attempt to find out how much historical data is actually needed or adequate, we monitor the values of the mean and variance of arrival time for all single-peak callers as more historical data (increased by day) are taken into computations. *We observe the convergence of means and variances.* As an example, Fig. 5 shows the convergence of mean and variance of arrival time of a single-peak caller as number of days towards the past increases.

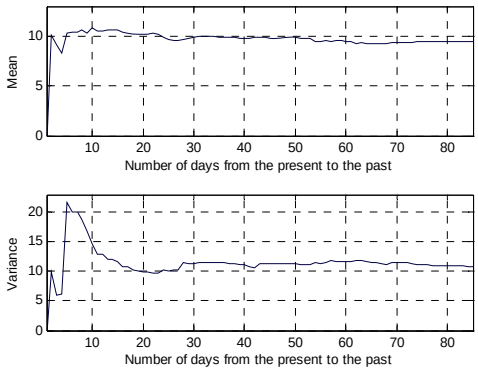

Figure 5: An example of observed convergence of mean and variance of arrival time of a single-peak caller.

It can be observed that the values of mean and variance converge to nearly constant after taking approximately the last 30 days of historical data. This means that the mean and variance of entire historical data are approximately the same as the mean and variance of the last 30 days of data. Since a single-peak caller is characterized by a normal distribution which depends on mean and variance, it implies that the last 30 days of data is adequate to capture the behavior of the single-peak caller. It is evident in Fig. 6 that the pdf from taking entire historical data and taking only last 30 days are similar.

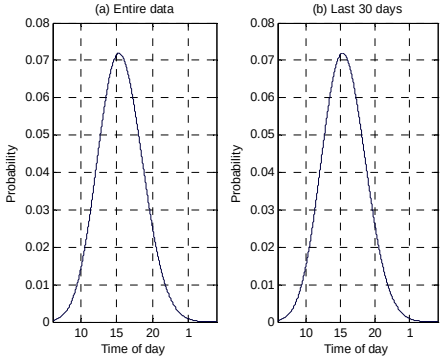

Figure 6: A comparison of pdf from (a) taking entire historical data and (b) taking only last 30 days of data.

The previous section also shows that the inter-arrival and talk time have exponential distribution *exp*(*m*) which depends only on the mean *m*. Therefore we examine the values of mean of inter-arrival and talk time as more historical data increases for all callers. However, we find that the convergence time is not observed.

A knowledge of mean and variance might not provide a pattern for a multi-peak caller due to the characteristics of the nonparametric density estimation. However, we believe that it captures physical behavior of a caller. In fact, the convergence of values of mean and variance of call arrival time of multi-peak callers is also observed. Figure 7 shows an example of a multi-peak caller whose mean and variance converge as the number of days towards the past increases. It can also be observed that the convergence time is approximately 60 days for this multi-peak caller.

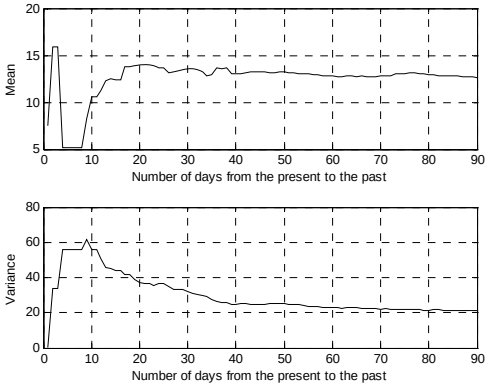

Figure 7: An example of observed convergence of mean and variance of arrival time of a multi-peak caller.

Figure 8 shows the pdf from taking entire historical data and taking the last 60 days of a multi-peak caller whose values of mean and variance are shown in Fig. 7. From Fig. 8, it appears that both pdf are slightly different in shape even though the mean and variance are nearly the same.

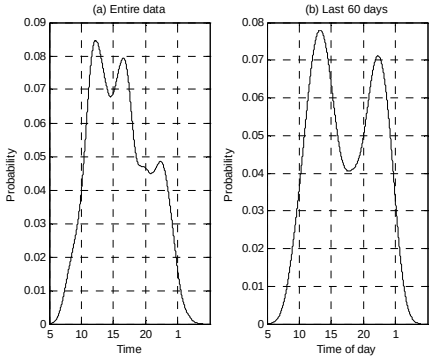

Figure 8: A comparison of pdf from (a) taking entire historical data and (b) taking only last 60 days of data.

We believe that the call logs represent human behavior associated with trends and changes of behavior over time. Considering historical data within the convergence time may provide us the recent trend of the data which can be more relevant to the future observation.

Our hypothesis is that the future behavior (pattern) of the caller based on call logs is more relevant to the pattern derived from the recent data (trend) than the pattern derived from the entire historical data (given that entire data are more than recent trend data). This hypothesis will be validated by the experiment conducted in the next section.

The crucial issue here is that of the convergence time (recent trend period) therefore we propose a simple technique for finding convergence time using a *Trace Distance* (*tD*).

Let us consider a sample of a converging signal shown in Fig. 9 where vertical axis represents amplitude and horizontal axis represents reversed time (time that runs towards the past) as similar to the plots shown Fig. 5 and 7.

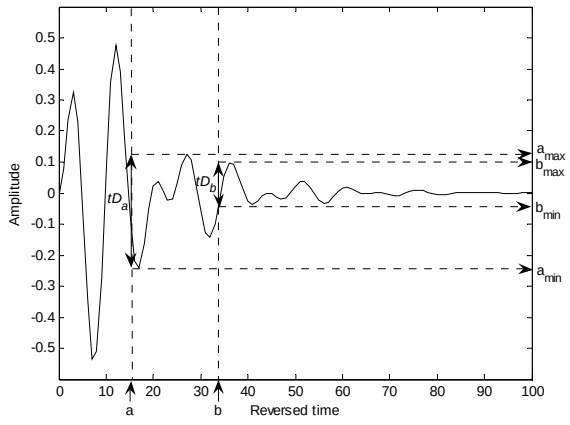

Figure 9: A converging signal which displays trace distances (tD_a and tD_b at reversed time *a* and *b* for demonstrating convergence time computation.

A trace distance at time k (tD_k) of signal s is a difference between the maximum amplitude and minimum amplitude from time k to infinity (most right-hand side of time k based on Fig. 9) which is given by Eq. (13).

$$tD_k = \| k_{\max} | - | K_{\min} \|, \quad (13)$$

where k_{max} and k_{min} are defined by Eq. (14) and Eq. (15) respectively.

$$k_{\max} = \max\{s(k), s(k+1), s(k+2), ..., s(\infty-1), s(\infty)\}, \quad (14)$$

$$k_{\min} = \min\{s(k), s(k+1), s(k+2), ..., s(\infty-1), s(\infty)\}. \quad (15)$$

Thus, the trace distances at time *a* and *b* shown in Fig. 12 can be computed as $tD_a = ||a_{max}| - |a_{min}||$ and $tD_b = ||b_{max}| - |b_{min}||$.

Therefore, the convergence time (*CT*) of the signal s is defined as the time that the trace distance (*tD*) reaches the predefined threshold (tD_{th}) as the trace distance computation starts from reversed time equals to zero to infinity which is given by Eq. (16).

$$CT_s = \{k \mid tD_k = tD_{th}, k \in \{0,1,2,..., \infty\}\}. \quad (16)$$

For our case, the signal s can be a reversed time series of mean and variance and the variable k represents the number of days towards the past.

An experiment is conducted to find convergence time of the callers in our datasets with tD_{th} set to 1. The convergence time is computed for each caller based on the arrival time. We find an interesting result of a relationship between the caller's convergence time and his/her number of peaks. *The result shows that as the number of peaks increases, the convergence time becomes larger*. Figure 10 shows a plot of the average convergence time versus the number of peaks.

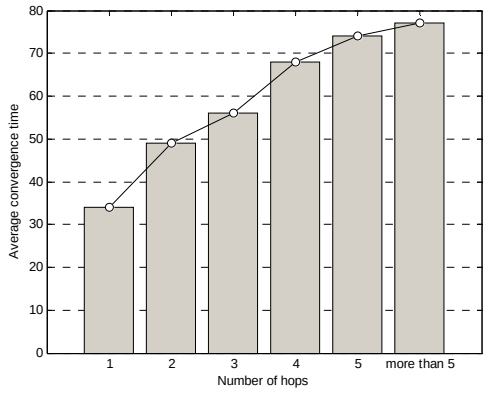

Figure 10: A plot of the number of peaks versus the average convergence time where the average convergence time becomes larger as the number of peaks increases.

We find that the result is reasonable. People who have random behaviors tend to not establish any behavioral pattern in a short period of time rather expand a recognizable structure over longer period of observation time. For example, a caller who was initially making lots of calls in the morning then started to make some calls in the evening and he/she eventually is making calls consistently in both morning and evening hours (two-peak caller).

It would take longer time to observe this caller's calling behavior than another caller who has been calling only during the morning hours (single-peak caller).

4. VALIDATION

To prove our hypothesis in the previous section that the future behavior (pattern) of the caller based on call logs is more relevant to the pattern derived from the recent data (trend) than the pattern derived from the entire historical data, we conduct an experiment.

The experiment is conducted to present the comparison of the relevance or similarity in caller behavior between the future observation and entire historical observation, and the similarity in caller behavior between the future observation and recent trend observation (convergence time).

To measure the similarity in calling behaviors, three measurements are chosen; *Correlation coefficient, Hellinger distance, and Relative entropy*. In addition, performance comparison of the Call Predictor (CP) proposed in [16] is also presented to observe the change in performance as the convergence time is considered.

Correlation coefficient [11] is a number between -1 and 1 which measures the degree to which two random variables are linearly related. A correlation coefficient of 1 implies that there is perfect linear relationship between the two random variables. A correlation coefficient of -1 implies that there is inversely proportional relationship between the two random variables. A correlation coefficient of zero implies that there is no linear relationship between the variables. In many applications, a correlation coefficient is used to measure how well trends in the predicted values follow trends in past actual values or how well the predicted values from a forecast model fit with the real-life data. A correlation coefficient (r) can be computed by Eq. (17) where P and Q are random variables which consist of small random variables $\{p(1), p(2), p(3), \ldots, p(N)\}$ and $\{q(1), q(2), q(3), \ldots, q(N)\}$ respectively.

$$r = \frac{\sum_{n=1}^{N} \left(p(n) - \overline{P}\right)\left(q(n) - \overline{Q}\right)}{\sqrt{\sum_{n==1}^{N} \left(p(n) - \overline{P}\right)^2 \left(q(n) - \overline{Q}\right)^2}}. \quad (17)$$

Hellinger distance ([8], [18]) has value between 0 and 1 which estimates the distance between probability measures. Let P and Q be the two probability measures which are N-tuple $\{p(1), p(2), p(3), \ldots, p(N)\}$ and $\{q(1), q(2), q(3), \ldots, q(N)\}$ respectively. P and Q satisfy $p_n \geq 0$, $\sum_n p_n = 1$, $q_n \geq 0$, and $\sum_n q_n = 1$. Hellinger distance is 0 implies that $P = Q$. Disjoint P and Q shows the maximum distance of 1. The Hellinger distance ($d_H^2(P,Q)$) between P and Q is given by Eq. (18).

$$d_H^2(P,Q) = \frac{1}{2}\sum_{n=1}^{N}\left(\sqrt{p(n)} - \sqrt{q(n)}\right)^2. \quad (18)$$

Relative entropy or Kullback Leibler distance [4] is a measure of the distance between two probability distributions. The relative entropy is a measure of the difference between assumed distribution Q and the true probability distribution P. Relative entropy is non-negative and is zero if $P = Q$. The relative entropy of Q from P is defined by (19) where P = $\{p(1), p(2), p(3), \ldots, p(N)\}$ and $Q = \{q(1), q(2), q(3), \ldots, q(N)\}$. Note that we use the convention that $0log(0/q) = 0$ and $plog(p/0) = 1$. The relative entropy ($D(P||Q)$) between P and Q can be computed by Eq. (19)

$$D(P \,\|\, Q) = \sum_{n=1}^{N} p(n) \log \frac{p(n)}{q(n)}. \quad (19)$$

In our case, P and Q are the N-tuple probability mass functions of the future observation and testing period respectively where the testing period can be either within the convergence time or entire historical data.

Phithakkitnukoon and Dantu [16] proposed a Call Predictor (CP) which computed receiving call probability and made the next-24-hour incoming call prediction based on caller's behavior and reciprocity. The caller's behavior was measured by the caller's call arrival time and inter-arrival time. The reciprocity was measured by the number of outgoing calls per incoming call and inter-arrival/departure time. The CP took into account the entire call historical.

In this experiment, we examine the performance of the CP with considering the convergence time of the call history and compare to the performance of the CP without considering the convergence time (or taking entire call history). The performance is measured in terms of *Error rate* which is defined as a ratio of the number of fault predictions to the total number of predictions made.

The experiment is conducted with 100 randomly selected callers including 30 single-peak callers and 70 multi-peak callers from our datasets. The most recent seven days of call logs are assumed to be future observation. The trace distance threshold tD_{th} is set to 1 to compute the convergence time (CT). The CP repeatedly computes the CT for each of the seven days prior to making call prediction.

Figure 11(a), 12(a), 13(a), and 14(a) show the comparisons of the computed correlation coefficients, Hellinger distance, relative entropy, and error rate of the CP respectively of all 100 callers between taking entire historical data (represented with an asterisk (*)) and taking data within the convergence time (represented with a circle (o)) where the first 30 callers are single-peak callers and at rest are multi-peak callers (31-100).

Figure 11(b), 12(b), 13(b), 14(b) show the changes in the values of correlation coefficient, Hellinger distance, relative entropy, and error rate of the CP respectively as the convergence time is considered.

It can be observed that the value of correlation coefficient increases as the convergence time is considered for all 100 callers which tells us that the recent caller behavior or calling pattern is more relevant (correlated) to the future calling pattern than the pattern observed from entire call history.

The values of Hellinger distance, relative entropy, and error rate of the CP decrease as the convergence time is considered which also confirms that the recent calling pattern is more relevant to the future pattern.

The experimental result is summarized in the Table 1 which lists the numerical average values of the correlation coefficient, Hellinger distance, relative entropy, and error rate of the CP when the entire data is considered, as well as when the data within the convergence time is considered, and their average changes for

categorized single-peak callers and multi-peak callers. Since the single-peak callers have normal distribution, the change in the similarity measures are relatively low compared to the multi-peak callers.

This experimental result shows that the data within convergence time is adequate to construct a predictive model and in fact it composes a recent pattern which is more similar or relevant to the future pattern than considering pattern composed by the entire historical data.

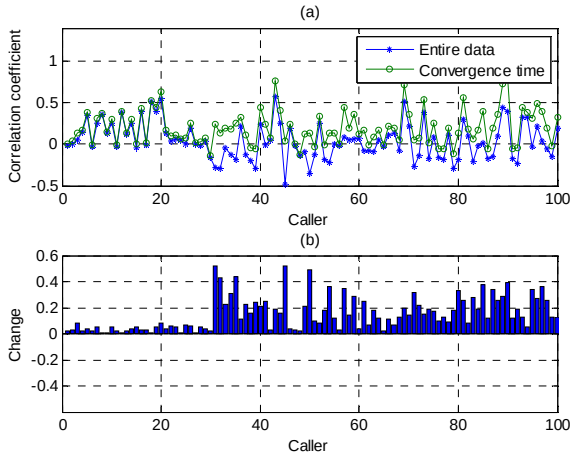

Figure 11: (a) Comparison of correlation coefficients and (b) its corresponding change from taking entire historical data to taking data within convergence time of each caller.

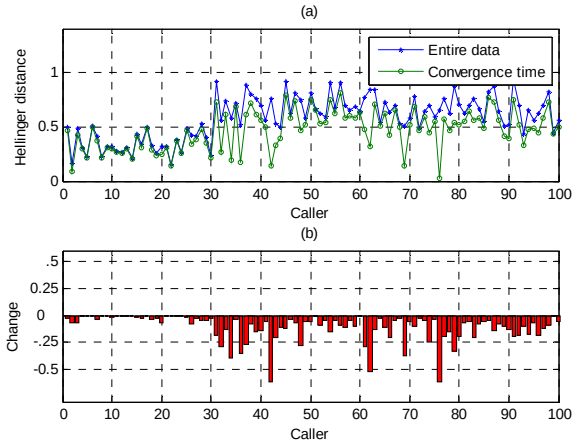

Figure 12: (a) Comparison of Hellinger distances and (b) its corresponding change from taking entire historical data to taking data within convergence time of each caller.

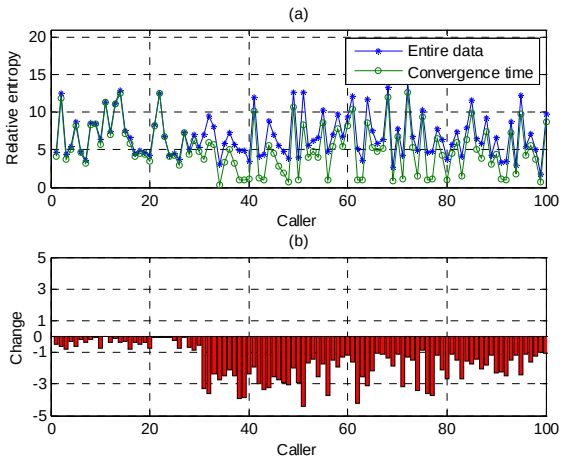

Figure 13: (a) Comparison of relative entropy and (b) its corresponding change from taking entire historical data to taking data within convergence time of each caller.

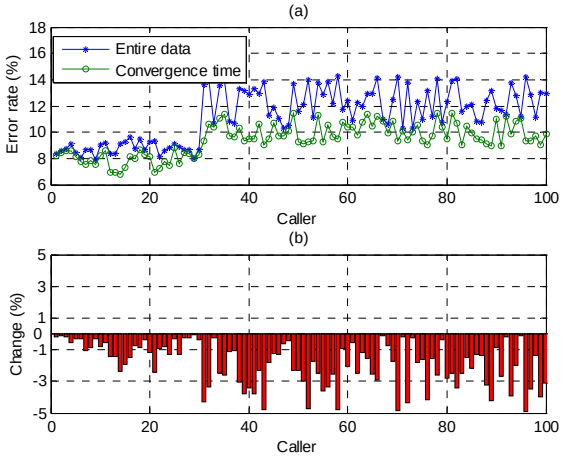

Figure 14: (a) Comparison of error rate of the call predictor and (b) its corresponding change from taking entire historical data to taking data within convergence time of each caller.

5. CONCLUSION

In this paper, we propose a technique to find the adequacy of historical call logs in order to capture the caller behavior (pattern). Firstly, the statistical analysis of real-life datasets to characterize caller behavior is carried out. We classify callers into two groups namely single-hop callers and multi-hop callers based on the distribution of the arrival time of the calls. We have verified the normal distribution for single-hop callers and estimated the distribution for multi-hop callers using kernel density estimator. We have also verified exponential distribution for inter-arrival time and talk time.

Since the caller behavior can be characterized by probability models which are used to predict or estimate the future behavior conditioned by a knowledge of the historical data, the question is how much historical data is adequate.

Table 1

The Average of Correlation coefficients (*r*), Hellinger distance (d_H^2), Relative entropy (*D*), and Error rate (*Err*) of taking entire historical data comparing to taking only data within the convergence time and its average change (increase(+) or decrease(-))

Callers	Average Measures of Taking Entire Data				Average Measures of Taking Data within Convergence Time				Average Change			
	r	d_H^2	***D***	***Err(%)***	***r***	d_H^2	***D***	***Err(%)***	***r***	d_H^2	***D***	***Err(%)***
(1-30) Single-peak	0.1476	0.6573	6.9377	8.751	0.1837	0.63	6.547	7.9153	+0.0361	-0.0273	-0.3907	-0.8357
(31-100) Multi-peak	0.0007	0.6791	6.8423	12.3672	0.2043	0.5329	4.6256	10.0429	+0.2036	-0.1462	-2.2167	-2.3243

Our study shows that the mean and variance of the arrival time converge to nearly constant as more historical data taken into account which means that only data within the convergence time is needed to construct a distribution model if the arrival time is characterized by mean and variance (which is true for normal distribution). The convergence is not observed for the inter-arrival and talk time however.

We also find that as the number of hops increases, the convergence time gets longer. Therefore we propose a simple technique to compute the convergence time using trace distance. In fact, the data within the convergence time is proven to be more relevant (or has higher correlation) to the future pattern of the caller by using correlation coefficient, Hellinger distance, and relative entropy. We believe that a call log is a human behavior related time series which is involved in trends or changes of behavior over time, therefore the mean and variance within the convergence time reflect the recent behavior or pattern of the caller.

We also show that our technique can be useful for constructing a predictive model for future incoming calls such as the Caller Predictor proposed by Phithakkitnukoon and Dantu [16] where its performance is improved by applying the proposed technique.

We will continue to investigate caller behavior and extend the idea of the adequacy of historical data to different types of data as our future direction.

6. ACKNOWLEDGMENT

Authors would like to thank Dr. Nathan Eagle and the Reality Mining project group at MIT for providing us the valuable dataset. This work is supported by the National Science Foundation under grants CNS-0627754, CNS-0619871, and CNS-0551694. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

7. REFERENCES

[1] A. Balachandran, G. M. Voelker, P. Hhl, and P. Venkat Rangan. Characterizing User Behavior and Network Perormance in a Public Wireless LAN. In *Proceedings of the ACM SIGMTRICS'02*, vol. 30, no. 1, 2002.

[2] M. Balazinska and P. Castro. Characterizing Mobility and Network Usage in a Corporate Wireless Local-Area Network. In *Proceedings of the 1st International Conference on Mobile Systems*, Applications, and Services, 2003.

[3] C. Chatfield. *The Analysis of Time Series An Introduction*. Chapman & Hall/CRC, 2004, pp. 1-9.

[4] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley, 1991, pp. 18.

[5] N. Eagle and A. Pentland. Social serendipity: Mobilizing social software. *IEEE Pervasive Computing*, vol. 4, no. 2, pp. 2834, 2005.

[6] N. Eagle, A. Pentland, and D. Lazer. Infering social network structure using mobile phone data. In *Proceedings of National Academy of Sciences*, 2006.

[7] N. Eagle. Machine Perception and Learning of Complex Social Systems. PhD Thesis, Massachusetts Institute of Technology, June 2005.

[8] M. Fannes and P. Spincemaille, The mutual affinity of randommeasures. In *eprint arXiv:math-ph/0112034*. December 2001.

[9] C. Guang, G. Jian, and D. Wei. Nonlinear-periodical networktraffic behavioral forecast based on seasonal neural network model. In *Proceedings of the International Conference on Communications, Circuits, And Systems (ICCCAS'04)*, vol. 1, pp. 683-687, 2004.

[10] M. Jones, J. S. Marron, and S. J. Sheather. A brief survey of bandwidth selection for density estimation. *Journal American Statistics Association*, 91, vol. 433,March 1996, pp. 401407.

[11] A. Leon-Garcia. *Probability and Random Processes for Electrical Engineering*, 2nd ed., Addison-Wesley, 1994, pp.135.

[12] Massachusetts Institute of Technology: Reality Mining Project. Available: http://reality.media.mit.edu/

[13] N. Moenne-Loccoz, F. Bremond, and M. Thonnat. Recurrent Bayesian Network for the Recognition of Human Behaviors from Video. In *Proceedings of the 3rd International Conference in Computer Vision Systems (ICVS'03)*, pp. 68-77, 2003.

[14] E. Parzen. On estimation of a probability density function and mode. *Annual Mathematic Statistics*, 33, vol. 3, 1962, pp. 10651076.

[15] A. Pentland and Andrew Liu. Modeling and Prediction of Human Behavior. In *Neural Computation*, vol. 11, pp. 229-242, 1999.

[16] S. Phithakkitnukoon and R. Dantu. Predicting Calls: New Service for an Intelligent Phone. In *Proceedings of the 10th IFIP/IEEE International Conference on Management of Multimedia and Mobile Networks and Services (MMNS'07)*, October 2007.

[17] P. Pirolli and W. Fu. SNIF-ACT: A Model of Information Foraging on the World Wide Web. In *Proceedings of the 9th International Conference on User Modeling*, pp. 4554, 2003.

[18] D. Pollard. Asymptopia. http://www.stat.yale.edu/pollard/ (Book in Progress), 1st edition, 2000.

[19] M. M. Rahman, K. Kakayama, and S. Ishikawa. Recognizing Human Behavior Using Universal Eigenspace. In *Proceedings of the 16th International Conference on Pattern Recognition (ICPR'02)*, vol. 1, pp. 10295, 2002.

[20] G. Resta and P. Santi. The QoSRWP Mobility and User Behavior Model for Public Area Wireless Networks. *ACM Computer Systems Organization*, Italy, Tech. Rep. 2006-TR-03, May 2006.

[21] S. J. Sheather and M. C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society*, Series B, 53, 1991, pp. 683690.

[22] W. Tych, D. J. Pedregal, P. C. Young, and J. Davies. An unobserved component model for multi-rate forecasting of telephone call demand: the design of a forecasting support system. *International Journal of Forecasting*, vol. 1, pp. 683-687, 2004.

[23] M. P. Wand and M. C. Jones. Multivariate plug-in bandwidth selection. *Computational Statistics*, 9, 1994, pp. 97117.