12-2011

Mobile Phone Graph Evolution: Findings, Model and Interpretation

Siyuan Liu

Lei Li

Christos Faloutsos

Lionel M. Ni

Mobile Phone Graph Evolution: Findings, Model and Interpretation

Siyuan Liu, iLab, Heinz College, Carnegie Mellon University
syliu@andrew.cmu.edu
Lei Li, Carnegie Mellon University
leili@andrew.cmu.edu
Christos Faloutsos, Carnegie Mellon University
christos@andrew.cmu.edu
Lionel M. Ni, Hong Kong University of Science and Technology
ni@cse.ust.hk

December, 2011

LARC-TR-02-11

LARC Technical Report Series: http://smu.edu.sg/centres/larc/larc-technical-reports-series/

ABSTRACT

What are the features of mobile phone graph along the time? How to model these features? What are the interpretation for the evolutional graph generation process? To answer the above challenging problems, we analyze a massive who-call-whom networks as long as a year, gathered from records of two large mobile phone communication networks both with 2 million users and 2 billion of calls. We examine the calling behavior distribution at multiple time scales (e.g. day, week, month and quarter), and find that the distribution is not only skewed with a heavy tail, but also changing at different time scales. How to model the changing behavior, and whether there exists a distribution fitting the multi-scale data well? In this paper, first, we define a δ stable distribution and a Multi-scale Distribution Fitting (MsDF) problem. Second, to analyze our observed distributions at different time scales, we propose a framework, *ScalePower*, which not only fits the multi-scale data distribution very well, but also works as a convolutional distribution mixture to explain the generation mechanism of the multi-scale distribution changing behavior. Third, *ScalePower* can conduct a fitting approximation from a small time scale data to a large time scale. Furthermore, we illustrate the interesting and appealing findings from our *ScalePower* model and large scale real life data sets.

Mobile Phone Graph Evolution: Findings, Model and Interpretation

Siyuan Liu #, Lei Li #, Christos Faloutsos #, Lionel M. Ni *
#*Carnegie Mellon University*
**Hong Kong University of Science and Technology*

***Abstract*—What are the features of mobile phone graph along the time? How to model these features? What are the interpretation for the evolutional graph generation process? To answer the above challenging problems, we analyze a massive who-call-whom networks as long as a year, gathered from records of two large mobile phone communication networks both with 2 million users and 2 billion of calls. We examine the calling behavior distribution at multiple time scales (e.g., day, week, month and quarter), and find that the distribution is not only skewed with a heavy tail, but also changing at different time scales. How to model the changing behavior, and whether there exists a distribution fitting the multi-scale data well? In this paper, first, we define a δ-stable distribution and a Multi-scale Distribution Fitting (MsDF) problem. Second, to analyze our observed distributions at different time scales, we propose a framework, *ScalePower*, which not only fits the multi-scale data distribution very well, but also works as a convolutional distribution mixture to explain the generation mechanism of the multi-scale distribution changing behavior. Third, *ScalePower* can conduct a fitting approximation from a small time scale data to a large time scale. Furthermore, we illustrate the interesting and appealing findings from our *ScalePower* model and large scale real life data sets.**

Categories and Subject Descriptors: H.2.8 Database applications: Data mining I.2.6 Artificial Intelligence: Learning - parameter learning

General Terms: Algorithms; Experimentation.

Keywords: Distribution; Generative Process; Lognormal; Convolution; Mobile Phone Graph.

I. Introduction

Mobile phone graph is attracting more and more attentions recently, and the feature study is a hot issue now [1], [3], [4], [8], [10], [14], [16]. One of the feature study is trying to figure out the patterns hidden in the graph. For example, the degree distribution of the nodes in the graph fits a heavy-tailed distribution. Moreover, heavy-tailed distribution is ubiquitous in an extremely wide range of phenomena, such as the heights of human beings, the degree of nodes in the Internet or the number of citations received by papers, which indicates that things always have a typical size or scale [6], [11], [18], [22], [23]. In our work, we study a large scale and appealing who-call-whom networks, mobile phone communication data, at multiple time scales to find the hidden surprising patterns which are not discovered and studied in current heavy-tailed distribution related work.

Given a very large amount of mobile phone communication records, what is the best way to summarize the multi-scale

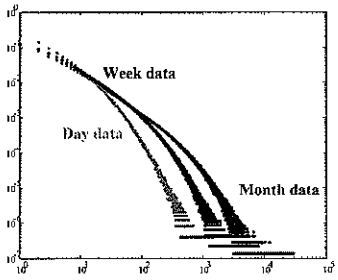
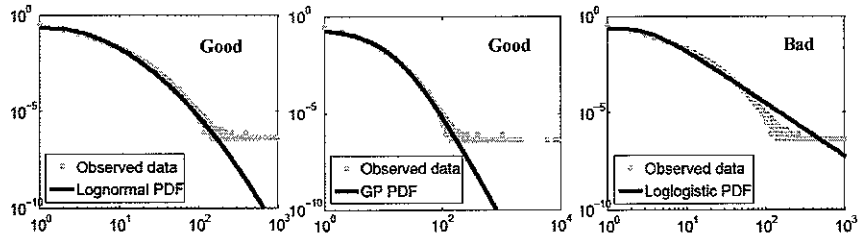


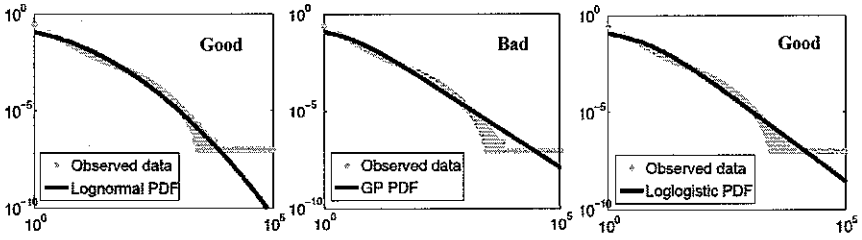
Fig. 1. Data features at multiple scales

(day, week, month and quarter) calling behavior of the data? Is the calling behavior in a day the same as the calling behavior in a week or a month? In the former study, a Lognormal distribution is proposed to fit the call duration in [13], a Double Pareto LogNormal distribution (DPLN) is proposed to fit the mobile phone call data in [25], while in [9] a Truncated Lazy Contractor (TLAC) model is designed to model the mobile phone call data. The above methods or findings may suit on a given certain time scale data set, but the problem is that how about the data features at multiple scales? We show the real data features by real call data at multiple scales in Figure 1. Note that the data distributions are different among day, week and month data. As the sequence problem, how about the data fitting at multiple scales? We illustrate this problem in Figure 2 where we fit day and bimonth data by Lognormal, Generalized Pareto (GP) and Loglogistic distributions. The observation is that GP fits day data well, while fits bimonth data bad; Loglogistic fits day data bad, while fits bimonth data well; promisingly, Lognormal fits both day data and bimonth data very well.

After examining large scale mobile phone communication records from two large cities in different countries, both with million order of magnitude mobile phones, billion order of magnitude call records, as long as one year, we more specifically analyze the number of calls, and have a surprising finding: the data distribution changes at different time scales. To investigate and interpret this finding, we first define a δ-stable distribution to describe an approximative stable distribution for multi-scale data fitting; second, we propose *ScalePower*, a framework which retrieves a distribution satisfying a δ-stable distribution, and then describes the distribution at multiple



(a) Day data



(b) Bimonth data

Fig. 2. Distribution fitting at multiple scales

scales. Moreover, *ScalePower* models and approximates the calling behavior by a novel convolutional generation. This convolutional distribution mixture describes the mobile phone uses as new-comer, leaver and consistent users at multiple scales. Last, we utilize two large mobile phone communication networks with millions of users and billions of calls to valid our model.

The contributions of this paper are below. First, we discover surprising patterns of mobile phone calling behavior at multiple scale. Second, we propose a δ-stable distribution as a family of distributions to best-fit multi-scale data. Third, we devise a framework, *ScalePower*, for best-fitting the mobile phone calling behavior at all observed scales. Fourth, we give the interpretation of the underlying calling behavior generation process.

The rest of the paper is organized as follows. In Section II, we provide a brief survey of other work that analyzed mobile phone records and probability distributions. In Section III, the preliminary information of our work is introduced. In Section IV, we formally define the problem. In Section V, we describe our proposed *ScalePower* framework and the hidden generation mechanism. We discuss our framework's goodness-of-fit at multiple scales and promising applications for our results in Section VI. In Section VII, we show the conclusions and future research directions.

II. Related work

Our work is related to three categories of current works. One is the mobile phone data fitting, one is the heavy-tailed distribution study, and the third one is the social graph mining. We briefly survey the related work as follows.

Mobile phone data fitting: Mobile phone data set is attracting more and more attentions recently. There are a number of current works giving insightful study of the mobile phone data set. Vaz de Melo et al. [9] investigated the surprising patterns for the call duration distribution of mobile phone users, and proposed a Trncated Lazy Contractor (TLAC) distribution to fit the call duration distribution, which is a truncated version of log-logistic distribution. Seshadri et al. [25] observed some distributions (of number of calls, distinct call partners, nd total talk time), and proposed a Double Pareto LogNormal distribution to fit the data. In [13], the authors proposed a log-normal distribution to fit the call duration. In [26], the authors found that the call duration neither exponentially nor log-normally distributed, and the distribution has a semi-heavy tail, which asks for a more heavy-tailed distribution. In our work, the major difference from theirs is that we study the distribution in a multi-scale scenario, that is, we study the distribution evolution at day, week, month and quarter time scales. Hence we not only best-fit the data at multi-scale, but also reveal the generation mechanism. Some unrevealed interesting patterns are also discovered in our work.

Heavy-tailed distribution: In probability theory, a random variable is said to be stable or have a stable distribution, if the random variable has the property that a linear combination of its two independent copies has the same distribution, up to location and scale parameters. The stable distribution family is sometimes referred to as the Lévy alpha-stable distribution [21]. The normal distribution is one family of stable distributions. Reed et al. [24], Clauset et al. [7], and Newman [22] studied the heavy tail distribution, and proposed the distribution function, approximation method and generation mechanism. Fofack et al. [11] and Nolan [23] studied the tail behavior, modes, modeling and accurate computation way of stable distribution. In our work, first, we propose a δ-stable distribution which describes a distribution stable at multiple scales. Second, the approximation of distribution for multi-scale data fitting is studied. These interesting works are not

discussed in the current literature.

Social graph mining: In recent years, social graph mining is a very hot topic. For example, Rodriguez et. al [12] tried to infer networks of diffusion and influence. Leskovec et. al [17], [19] studied graph over time by densification laws, shrinking diameters and possible explanations, and provided a graph generator based on a forest fire spreading process to study the graph evolution. McGlohon et. al [20] studied patterns in weighted graphs and proposed a generator. In our work, we investigate the graph generated from mobile phone networks at multiple scales, and propose the fitting method.

III. PRELIMINARIES

In this section, first, we describe the data sets we study; second, we introduce the related distributions and tests.

A. Mobile phone data sets at multiple scales

In our work, we study two large scale mobile phone data sets. *Data set 1* is collected from a city in Asia. The size of the city is around 8700 km^2. In this city, there are four million mobile phones and more than ten million records per day. The size of the raw data set that we collected from 1^{st} January, 2008 to 31^{st} December, 2008, is around 0.7 Terabytes. *Data set 2* is half-year collected from a private mobile phone company of a large city, with more than three million users and one billion phone call records, spanning 0.1 Terabytes.

The data sets are both generated from the *Call Detail Record* (CDR) which is the information related to mobile phone communication, such as caller ID, callee ID, call start time and call duration. In the following study, to make our method and findings clear, we illustrate our work by one attribute of the mobile phone communication data, that is the number of calls. The number of calls is defined as the total number of calls per user in a given time interval. The number of calls distribution is the the data distribution of all the users' call in a given time period (time scale). Time scale is defined as a time period that we observe the accumulated data. For example, a day means we observe the data by one day as the time unit.

In this paper, we emphasize that our interest is in aggregating statistical analysis and therefore, we do not study any particular individual's calling behavior. In order to preserve the user privacy and anonymity, data that could identify users, e.g., the phone numbers, is not utilized in this study. We take *Data set 1* as our study data set, and *Data set 2* is utilized to confirm the same findings. Without specification, we utilize *Data set 1* by default.

B. Distribution fitting

1) Heavy-tailed distribution: In probability theory, a heavy-tailed distribution has a much heavier tail (not exponentially bounded) than an exponential distribution [21]. In the heavy-tailed distribution family, there are left-tailed, right-tailed, two heavy tailed distributions. A random variable X with a distribution function F is said to have a heavy right tail if, for all $\lambda > 0$,

$$\lim_{x\to\infty} e^{\lambda x} \Pr[X > x] = \infty \tag{1}$$

The definitions of heavy-tailed for left-tailed or two tailed distributions are similar.

In heavy-tailed distribution family, there are several one tailed distributions, such as Lognormal (LN), Loglogistic (LG) and Generalized Pareto (GP) distributions. Accordingly, there are some mixture forms of above distributions, such as Double Pareto Lognormal (DPLN) distribution and Pareto Lognormal distribution [23]. For example, a Lognormal distribution is a probability distribution of a random variable whose logarithm is normally distributed. The parameters of a Lognormal distribution are denoted μ and σ, which are the mean and standard deviation, respectively. The probability density function (PDF) of a Lognormal distribution is as follows.

$$f_X(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}} \tag{2}$$

where $x > 0$.

The cumulative distribution function (CDF) of a Lognormal distribution is

$$F_X(x;\mu,\sigma) = \frac{1}{2} erfc[-\frac{\ln x-\mu}{\sigma\sqrt{2}}] = \Phi(\frac{\ln x-\mu}{\sigma}) \tag{3}$$

where erfc is the complementary error function, and Φ is the standard normal CDF.

To fit a given data set with a heavy-tailed distribution, a PDF or a CDF are computed, and then we fit the empirical CDF or PDF with the observed data distribution. In our work, we fit the given data by LN, LG, GP and DPLN distributions.

2) Goodness-of-fit: For testing a hypothesis whether a distribution function fits a given data, there are Kolmogorov-Smirnov (KS) test, Berk-Jones test, score test and their integrated versions [5], [15]. Kolmogorov-Smirnov score quantifies a distance between the empirical distribution function (ECDF) of the sample and the cumulative distribution function of the reference distribution, or between the empirical distribution functions of the two samples, as a sequence, it can be used to decide whether a sample comes from a population with a specific distribution. Kolmogorov-Smirnov score is defined as follows.

The empirical distribution function F_n for n independent and identically distributed random variable X_i is defined as

$$F_x(x) = \frac{1}{n}\sum_{i=1}^{n} I_{X_i\leq x} \tag{4}$$

where $I_{X_i\leq x}$ is an indicator function(if $X_i \leq x$, it equals 1; 0, otherwise).

Kolmogorov-Smirnov score for a given cumulative distribution function $F(x)$ is

$$S_x = \sup_x |F_n(x) - F(x)| \tag{5}$$

where sup x is the supremum of the distance set.

As Kolmogorov-Smirnov test has a long tradition in statistics, can be a goodness-of-fit test for any statistical distribution and there are no other tests which clearly perform better, hence in our work, we evaluate the goodness-of-fit by the KS test [15]. We say a distribution fit the data good at multiple scales, if this distribution fits the data good at each scale.

IV. Problem: *MsDF*

In this section, we formally define the problem we try to solve in our work, that is, Multi-scale Distribution Fitting (MsDF) problem.

A. Definition

Recall the definition of Stable distribution, we introduce the following definition of δ-stable distribution.

Definition 1: (δ-stable distribution) A family of distribution $D(\theta)$ is said to be δ-stable, if two independent random variables $(X, Y) \sim D(\theta)$, there exist θ_o, such that $Z = X + Y$ can be approximated by a distribution $D(\theta_o)$, where

$$\max |D(\theta_o) - D(X + Y)| < \delta \tag{6}$$

θ, θ_o and δ are parameters.

The intuition of Definition 1 is that the PDF convolution can be approximated by a distribution from the same distribution family. For example, Lognormal distribution is a δ-stable distribution. A sum of Lognormal distribution can be approximated by a Lognormal distribution [2], [27], [28]. Let a random variable X=ln$Y$, then

$$W = \sum_{i=1}^{N} Y_i = e^{X_1} + e^{X_2} + \cdots + e^{X_N} \approx e^{Z} \tag{7}$$

where the random variable Z possesses a normal distribution.

In Figure 3, we utilize a simulation result to illustrate such a characteristic of the Lognormal distribution. In the figure, the black line is the actual Lognormal sum of two Lognormal distributions, and the red line is a Lognormal distribution. The observation is that the grey line can approximately fit the black line. The multiple scale fitting can be intuitively modeled by a convolution. For example, a week data is a sum of seven days data. In our work, the score of goodness-of-fit is KS score. The problem we try to resolve is formally defined in the following subsection.

B. MsDF problem

In our work, we investigate the data at multiple scales, and want to find a distribution can fit the data at all observed data scale. The formal definition of this problem is below.

Multi-scale Distribution Fitting (MsDF) problem: Given a data at multiple scales, how to find a distribution satisfying δ-stable distribution.

To solve MsDF problem, we have to solve two problems. Problem 1: how to find a good distribution fitting at a given scale. Problem 2: how to find a δ-stable distribution fitting the given data at multiple scales. In our work, specifically, for the number of calls distribution fitting, we try to retrieve a distribution satisfying δ-stable distribution, that is, the KS score is less than a given parameter δ at multiple scales.

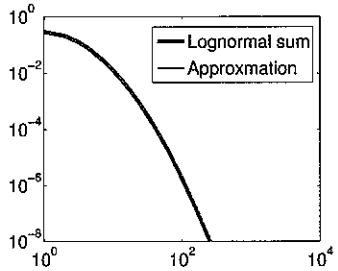


Fig. 3. Lognormal sum

TABLE I
GOODNESS-OF-FIT RESULTS (KS SCORE)

	1	7	30	60	90
LN	0.1181	0.1273	0.1291	0.1465	0.1553
DPLN	0.3063	0.2987	0.2570	0.2780	0.2830
LG	0.3685	0.2808	0.2169	0.2647	0.2154
GP	0.2251	0.1511	0.1483	0.1665	0.1755

V. Proposed Framework: *ScalePower*

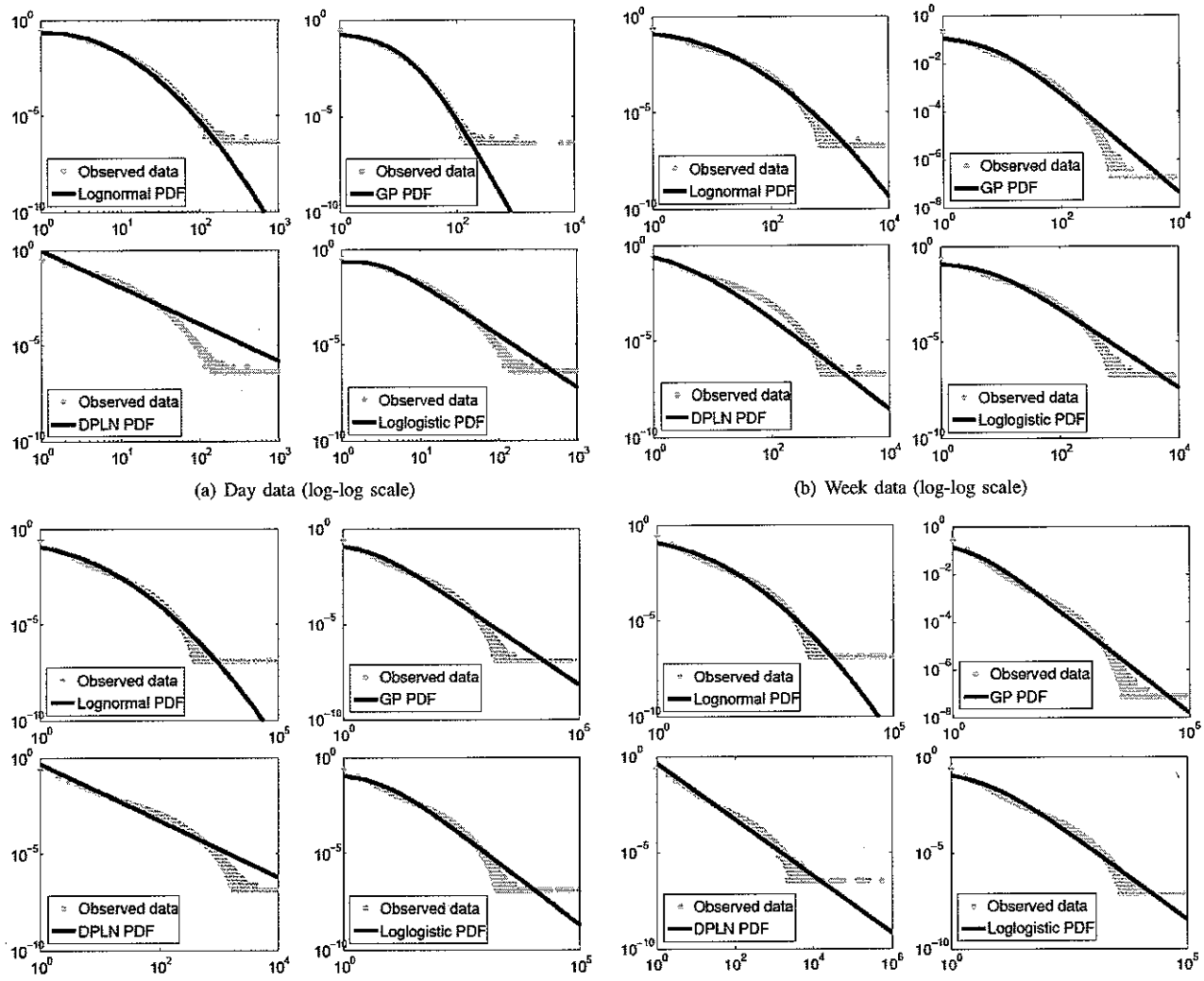
In this section, we introduce our proposed framework, *ScalePower*. First, we conduct a goodness-of-fit at multiple scales to find out a δ-stable distribution, and second, we propose a convolutional mixture distribution to approximate the δ-stable distribution.

A. Goodness-of-fit

In the goodness-of-fit, we conduct KS test on the distribution fittings at multiple scales, and find out the distribution satisfying δ-stable distribution. In our work, δ is set as 0.16, which is considered as a good fitting result test [15], [21]. The test results of the goodness-of-fit are reported in Figure 4 and Table I, and the independence study of the number of calls is shown in Figure 5.

In Figure 4, we show the distribution fitting results at multiple scales and the fitting is conducted on the PDF. In the figure, the grey points are the observed data at multiple scales, and the black line is the fitted result by a probability distribution. There are four data scales in our test, that is, day data, week data, month data and quarter data (i.e., three months data). There are four distributions in our test, that is, Lognormal distribution, GP distribution, Loglogistic distribution and DPLN distribution. At day scale test, Lognormal distribution and GP distribution show a best-fit of the data. At week scale test, Lognormal distribution is better than the other three distributions, and GP distribution is similar to Loglodistic distribution. The similar observation can be found at month scale data and quarter scale data. In a conclusion, Lognormal fitting is the best considering all the observed data scales. To better investigate the goodness-of-fit of the four distributions at multiple scale data, we conduct KS test, and the results are reported in Table I.

In Table I, we report the goodness-of-fit results. In the table, the first row are the number of days indicating the time scale that we study, and the first column are the distributions we



(a) Day data (log-log scale)

(b) Week data (log-log scale)

(c) Month data (log-log scale)

(d) Quarter data (log-log scale)

Fig. 4. Distribution fitting results at multiple scales

investigate. The goodness-of-fit results confirm that Lognormal fits all the data at different scales. Hence we can conclude that the number of calls can be fitted by a Lognormal distribution.

In Figure 5, the scatter plot of number of calls at multiple scales are illustrated. X-axis and Y-axis are the data at the same scale. In Figure 5 (a), we plot the day data against another day data. In Figure 5 (b), we plot the month data against another month data. The observation is that the data are scattered in the figure, which means the calling behavior is independent. Hence in the following sections, we assume the calling is independent, and model a user's number of calls as an iid random variable.

B. Convolutional mixture

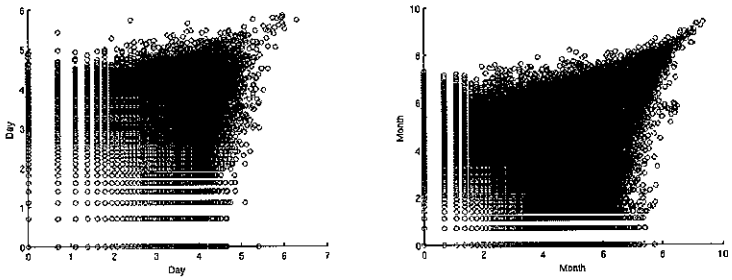
In practice, the mobile phone graph data can be incrementally summed. For example, two day data equal to the sum of two individual data. This intuition triggers us to design a convolutional mixture model to describe the generation mechanism of mobile phone graph evolution.

ScalePower not only fit the data at the single scale and multiple scales, but also interpret the generation mechanism of Lognormal-fitting-well at multiple scales. The basic idea is approximating a convolution of two Lognormal random variables by one Lognormal distribution.

Theorem 1: Given two random variables X and Y, we assume X follows a Lognormal distribution, and Y follows a Lognormal distribution, then the convolutional result of X and Y, $(X + Y)$, can be approximated by one Lognormal distribution [2].

Theorem 1 actually can be extended into a number of Lognormal distributions. In [27], [28], the techniques are developed to approximate the sum of Lognormals.

Given two independent random variables X (e.g., the number of calls of a set of persons in the 1^{st} month) and Y (e.g., the number of calls of a set of persons in the 2^{rd} month), assume that X follows a Lognormal distribution, and



(a) Day vs. day (log-log scale) (b) Month vs. month (log-log scale)

Fig. 5. Scatter plot at multiple scales

Y follows a Lognormal distribution. $\widetilde{X}$ is the augmented X, and $\widetilde{Y}$ is the augmented Y. The physical meaning of augmented variable here is that not only the number of calls of a set of persons in one month, but also plus the the number of calls of the persons who make calls in the other month but not in this month.

$\widetilde{X}$ follows the following probability distribution,

$$\widetilde{X} = \begin{cases} 0, & p \\ LN(\mu_x, \sigma_x), & (1-p) \end{cases} \tag{8}$$

Equation 8 means $\widetilde{X}$ equals 0 with a probability as p, and equals $LN(\mu_x, \sigma_x)$ with a probability as $(1-p)$.

$\widetilde{Y}$ follows the following probability distribution,

$$\widetilde{Y} = \begin{cases} 0, & q \\ LN(\mu_y, \sigma_y), & (1-q) \end{cases} \tag{9}$$

So $\widetilde{Y}$ equals 0 with a probability as q, and equals $LN(\mu_y, \sigma_y)$ with a probability as $(1-q)$.

Let Z=X+Y, and then $\widetilde{Z}$=$\widetilde{X}$+$\widetilde{Y}$. The convolutional probability distribution of ($\widetilde{X}$+$\widetilde{Y}$) is $P(\widetilde{Z})$.

Theorem 2: P follows the following probability distribution,

$$P(\widetilde{Z}) = \begin{cases} 0, & pq \\ LN(\mu_x, \sigma_x), & p \\ LN(\mu_y, \sigma_y), & q \\ LN(\mu_x, \sigma_x) + LN(\mu_y, \sigma_y), & (1-p)(1-q) \end{cases} \tag{10}$$

and can be approximated by a Lognormal distribution.

C. Parameter estimation and prediction

The above theorems tells what the bi-month distribution will look like based on one month's distribution. Our framework provides a prediction model for the number of call at a larger scale. Here is the basic flow of our prediction. We will use monthly data as an example.

Step 1 estimate the μ_1, σ_1 with LN($x; \mu_1, \sigma_1$).

Step 2 estimate μ_0, σ_0 with LN($y; \mu_0, \sigma_0$), where y is the number of call for those who made call on 1st month but not second month.

Step 3 LN(μ_3, σ_3) $\leftarrow$ fitting the convolution of LN($x; \mu_1, \sigma_1$) by matching the moments of the Lognormal and the convolution.

Step 4 sample $z_{1...N}$ from the distribution as described in Eq. (10).

Step 5 estimate the μ_z, σ_z with LN($z; \mu_z, \sigma_z$).

Using the above procedure we can effectively estimate the distribution of bi-monthly number of call from one-month data (similarly for daily, weekly and other scales). We will show such our method generates good fits in the following experiment section.

VI. Experimental Results

In this section, we utilize our large scale data sets to valid our proposed framework, *ScalePower*, and discuss the potential applications of *ScalePower*.

A. ScalePower *validation*

In *ScalePower* validation, first, we valid *ScalePower* as a convolutional mixture of Lognormal approximation; second, we utilize *ScalePower* to approximate large scale data from a fitted small scale data. Third, we study and interpret the parameters in *ScalePower*. In Figure 6, we utilize real data to illustrate the Lognormal sum approximation result. The grey points are the real data, the black line is the Lognormal sum result, and the red line one Lognormal which approximates the Lognormal sum. The experiment result shows that the approximated Lognormal fits the real data and Lognormal sum very well. Hence, *ScalePower* can utilize a small scale data fitting result (a Lognormal distribution) to approximate a large scale data fitting. In Table II, we report the fitted Lognormal parameters at multiple scales. The first row are the number of days (scale), and the first column are the two parameters of Lognormal. The result shows that as the time scale becomes larger, μ and σ are becoming much larger, which means that the mean and variance of the data are becoming larger. In practice, as the observed calling behavior data scale becoming longer, the mean of the number of calls become larger and the variance also become larger. Based on the parameters in Table II, we can utilize *ScalePower* to approximate the larger time scale data distribution. The experiment results are reported in Table III.

In Table III, the approximated result is reported. In the table, the first row are the prediction time scales, for example, two days mean that we utilize a day data to predict two days data.

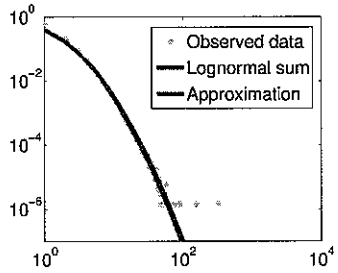


Fig. 6. Lognormal sum approximation

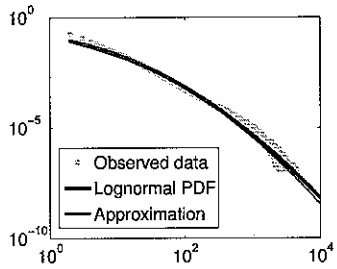


Fig. 7. Approximation and fitting in *Data set 2*

TABLE II
LOGNORMAL FITTING PARAMETERS

	1	7	30	60	90
μ	1.0578	1.9144	2.1186	2.1256	2.1378
σ	0.9691	1.5549	1.8740	2.0155	2.1081

TABLE IV
ScalePower PARAMETERS AT MULTIPLE SCALES

	Day	Week	Month
p	0.2154	0.2132	0.2206
q	0.2437	0.2441	0.2401
$1-p-q$	0.5409	0.5427	0.5393

The first column are the estimated parameters (μ_e and σ_e) and predicted parameters (μ_p and σ_p). The result shows that our method results in a very promising prediction accuracy. At multiple scales, our method all work well.

In Table IV, we report the parameters of *ScalePower* at multiple scales. In the table, the first row are the time scales, and the first column are the parameters of *ScalePower* (p, q and $1-p-q$). p indicates the percentage of new-comers in the networks, q indicates the leavers' percentage, and $(1-p-q)$ indicates the percentage of consistent users in the networks. An interesting observation is that the percentages of three types of users in the networks are stable no matter how long the observed time is, even though the individuals of a category may be not the same.

In Figure 7, we report the month data fitting in *Data set 2*. In the figure, the grey points are the observed one month data, the black line is the Lognormal distribution fitting result, and the red line is the approximation result by a Lognormal distribution based on our *ScalePower* method. The result shows that Lognormal distribution fits the data very well, and the approximation result is very promising, which is close to the estimated Lognormal distribution from the real data. The same results can be checked at multiple scales.

B. *Potential applications of* ScalePower

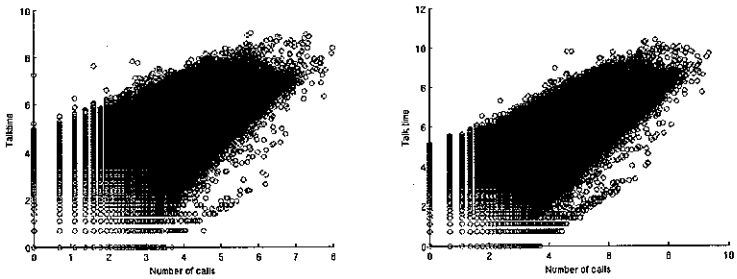
Thus far, we introduce our *ScalePower*, and valid it at multi-scale data. The approximation and fitting results are very promising. While several applications are possible, we focus on three in particular. First, the outlier detection in the calling behavior. Second, the social ability study. Third, distribution prediction.

TABLE III
APPROXIMATED LOGNORMAL PARAMETERS

	Two days	Two weeks	Two months
μ_e	1.3269	2.0132	2.1256
σ_e	0.9753	1.6013	2.0155
μ_p	1.3310	2.0014	2.1148
σ_p	0.9513	1.5872	2.0027

1) Outlier detection: In Figure 4, for the Lognormal fitting results, we can find out that there are two categories of outliers identified by the fitting line. First, a category with extreme limited number of calls in a time period. For example, in a month, there are more than 25 % of the total users calling once. An interesting finding is that this percentage is close to the leaver's percentage in our *ScalePower* model, which means that these one time calling persons may leave our observed networks soon. Second, a category with extreme high number of calls in a time period. For example, in a month, there are nearly 0.05 % of the total users calling ten thousand times. When we survey and check the utilization of these numbers, surprisingly, we find out that these numbers are a kind of service agency, which has extreme high call volumes.

2) Social ability study: In Figure 8, we illustrate the scatter plot of number of calls and talk time at multiple scales. In the figure, X-axis is the number of calls, and Y-axis is the talk time. Talk time is defined as the total call duration (second) in a given time period. One observation is that the scatter plot can be divided into two parts. The first one is a well scattered part at the left-bottom of the figure. The second one is the left-off of the figure, which centers on the diagonal of the figure. The interpretation of such a pattern is as follows. The first part means if a person gives a small number of calls, this person's talk time varies from short time to long time, which means an unstable social ability. The persons in this category may be random calling, traveler or service agency in the networks. The second part means that if a person gives a larger number of calls, this person's talk time becomes larger, respectively, which means a stable social ability. Interestingly, the percentage of the persons in this category is close to the consistent users' percentage in our *ScalePower* model. These persons not only have stable social ability, but also consistent in the networks.



(a) Week (log-log scale) (b) Month (log-log scale)

Fig. 8. Scatter plot of number of calls and talk time at multiple scales

3) Distribution prediction: As discussed in Figure 6, Table III and Figure 7, our *ScalePower* model can approximate a large scale data from a small scale data. For example, *ScalePower* can approximate a data set with two days scale from a data set with one day scale, which means *ScalePower* can conduct a prediction on the distribution at multiple scales. This is a promising application of *ScalePower*, which needs more study and investigation.

VII. Conclusion and Future Work

In this paper, we explored user calling behaviors in the large mobile phone communication data, specifically we analyzed millions of users and billions of phone call records at multiple scales (data, week and month data). The main contributions of this paper are:

- Discovery of surprising patterns of number of calls at multiple time scales.
- Identification of a δ-stable distribution to fit multi-scale data.
- Proposal of *ScalePower* model to conduct fitting and approximation at multiple time scales.
- Study of generation mechanism to explain the surprising patterns and fitting results.

Future work could focus on how to utilize our model to other data distributions of mobile phone networks, e.g., call duration and number of friends. A second promising direction is to spur further studies involving other data sets and underlying generative processes at multiple scales.

References

[1] L. A. Adamic and B. A. Huberman. The web's hidden order. *Commun. ACM*, 44:55–60, September 2001.

[2] N. Beaulieu and Q. Xie. An optimal lognormal approximation to lognormal sum distributions. *IEEE Trans. on Vehic. Tech.*, 53(2):479 – 489, 2004.

[3] Z. Bi, C. Faloutsos, and F. Korn. The "dgx" distribution for mining massive, skewed data. In *Pro. of ACM SIGKDD 2001.*

[4] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. *Comput. Netw.*, 33, June 2000.

[5] M. C. Bryson. Heavy-tailed distributions: Properties and tests. *Technometrics*, 16(1), 1974.

[6] D. Chakrabarti and C. Faloutsos. Graph mining: Laws, generators, and algorithms. *ACM Comput. Surv.*, 38, June 2006.

[7] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Rev.*, 51, November 2009.

[8] C. Cortes, D. Pregibon, and C. Volinsky. Communities of interest. In *Proc. of IDA 2001.*

[9] P. O. S. V. De Melo, L. Akoglu, C. Faloutsos, and A. A. F. Loureiro. Surprising patterns for the call duration distribution of mobile phone users. In *Proc. of ECML PKDD 2010.*

[10] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *Proc. of ACM SIGCOMM 1999.*

[11] H. Fofack and J. Nolan. Tail behavior, modes and other characteristics of stable distributions. *Extremes*, 2:39–58, 1999.

[12] M. Gomez Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. In *Proc. of ACM SIGKDD 2010.*

[13] J. Guo, F. Liu, and Z. Zhu. Estimate the call duration distribution parameters in gsm system based on k-l divergence method. In *Proc. of IEEE WiCOM 2007.*

[14] S. Keshav. Why cell phones will dominate the future internet. *SIGCOMM Comput. Commun. Rev.*, 35, April 2005.

[15] A. Koning and L. Peng. Goodness-of-fit tests for a heavy tailed distribution. Econometric institute report, Erasmus University Rotterdam, Econometric Institute, Nov 2005.

[16] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Extracting large-scale knowledge bases from the web. In *Proc. of VLDB 1999.*

[17] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proc. of ACM SIGKDD 2005.*

[18] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data*, 1, March 2007.

[19] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data*, 1, March 2007.

[20] M. McGlohon, L. Akoglu, and C. Faloutsos. Weighted graphs and disconnected components: patterns and a generator. In *Proc. of ACM SIGKDD 2008.*

[21] D. Montgomery and G. Runger. *Applied Statistics and Probability for Engineers*. Wiley India Pvt. Ltd., 2007.

[22] M. E. J. Newman. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46, May 2007.

[23] J. P. Nolan. *Stable Distributions - Models for Heavy Tailed Data.* Birkhauser, Boston, 2011.

[24] W. J. Reed and M. Jorgensen. The double pareto-lognormal distributionła new parametric model for size distributions. *Communications in Statistics - Theory and Methods*, 33, 2004.

[25] M. Seshadri, S. Machiraju, A. Sridharan, J. Bolot, C. Faloutsos, and J. Leskove. Mobile call graphs: beyond power-law and lognormal distributions. In *Proc. of ACM KDD 2008.*

[26] D. Willkomm, S. Machiraju, J. Bolot, and A. Wolisz. Primary users in cellular networks: A large-scale measurement study. In *Proc. of IEEE DySPAN 2008.*

[27] J. Wu, N. Mehta, and J. Zhang. Flexible lognormal sum approximation method. In *Proc. of IEEE GLOBECOM 2005.*

[28] Z. Wu, X. Li, R. Husnay, V. Chakravarthy, B. Wang, and Z. Wu. A novel highly accurate log skew normal approximation method to lognormal sum distributions. In *Proc. of IEEE WCNC 2009.*