

Mobile Phone Social Networks: Beyond Power-Law and Lognormal Distributions

Mukund Seshadri
XXXXXX Applied Research
XXXXXX, California, USA
mukund.seshadri@xxxxxx.com

Sridhar Machiraju
XXXXXX Applied Research
XXXXXX, California, USA
sridhar.machiraju@xxxxxx.com

Ashwin Sridharan
XXXXXX Applied Research
XXXXXX, California, USA
ashwin.sridharan@xxxxxx.com

Jean Bolot
XXXXXX Applied Research
XXXXXX, California, USA
bolot@xxxxxx.com

Christos Faloutsos
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
christos@cs.cmu.edu

Jure Leskovec
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
jure@cs.cmu.edu

ABSTRACT

We analyze a massive social network gathered from a large mobile phone operator's records, comprised of millions of users and tens of millions of calls. We examine the following questions: what is the distribution of phone calls per customer; total talk time per customer; and distinct partners per customer? We find that these distributions are skewed, and that they significantly deviate from what would be expected by conventional wisdom, namely power-law and lognormal distributions.

Instead, we propose the use of the recently formulated DPLN (Double Pareto Log Normal) distribution, and find a good fit for all metrics of interest (number of phone calls, total talk time, etc). Furthermore, we describe a generative process based on "social wealth", that naturally results in the DPLN distributions we observe. Finally, we study the evolution of the parameters of these distributions over time, and their significance in modeling and forecasting.

Categories and Subject Descriptors

TODO:Graph Mining [Social Networks]: Clustering

General Terms

TODO:DPLN, SVD

Keywords

TODO:Heavy tailed distributions

1. INTRODUCTION

[1.25 pages including abstract and title]

Currently Empty.

2. SKEWED DISTRIBUTIONS: BACKGROUND AND SURVEY

[1.5 Columns]

Background on Power Laws and Lognormals: Math definitions and Survey of applications/discoveries in Real-world Data Mining Literature.

[TODO:Outline below of different pieces of related work - probably will have to change]:

Numerous data sets have been studied and hypothesized to follow heavy-tailed distributions. Typically the researchers have shown a good fit to a power-law distribution. In several cases only a part of the data is fit, for example, the tail.

In many cases, the lognormal distribution has also been fitted to such datasets. In fact, it is hard to distinguish between power-law and lognormal when considering only the tail of the observed distribution. The proposed generating mechanisms for the two distributions are also intrinsically similar. Therefore there is a controversy over which is the better fit...

We quickly list some of the important studies in this area, in chronological order.

- 1.City Sizes [Gibrat?]
- 2.Income[Champernowne?]
- 3.File Size Distribution[Mitzenmacher?]
- 4.WWW page hyperlink degrees[Barabasi?]
- 5.AS degree distribution[Faloutsos?]
- 6.Barabasi Telecom work?
- 7.IBM phone graph?
- 8.[Any more???

Currently empty.

TODO: Maybe end section with Deficiencies in Related Work and how we plan to address it, leading naturally to DPLN...

3. DPLN: A PRIMER

Empty File

4. OUR DATASET DESCRIPTION

1-1/3 pages

1/3 column: General: What is PCMD, how is it collected, what are the general fields, what is the general size.

1/3 column: Specific: What are our metrics of interest, and fields of interest.

1/3 column: Table indicating different volumes and numbers.

1/3 column: Caveats and incompleteness.

Empty File

5. ANALYSIS OF DISTRIBUTION

[3.5 pages total]

We now present our analysis of the collected call records and associated social graph. The focus of our analysis is on the distribution of various social and calling behaviors of customers (of the mobile phone operator). [TODO: **Why focus on distributions??**]. We first present the distribution of our three main metrics of interest, per user, extracted from all call records corresponding to one month and one geographic area (one switch, to be exact). Then, we analyze statistical distributions to model our observed data, and present our results on the best fit. Finally, we will demonstrate that the general shape of the distribution remains the same across other time periods and other switches.

5.1 General Shape of Distribution

[0.75 page: 1 column - 3 graphs: Total no. of friends, calls and duration; 1/2 column text...]

Consider Figures 1, 2 and 3. These figures display the distribution of three per-user metrics respectively: number of call-partners, number of calls, and total call duration (in seconds). These metrics are the total accumulated over one month (mid-May to mid-June 2007)[**TODO exact dates**] by each customer. Each figure is a histogram in the log-log scale, with the X-axis representing the range values of the metric under consideration, and the Y-axis representing the number of users with corresponding metric values.

These figures show that all three metrics have the same general shape. The shape corresponds to a heavy-tailed distribution, since the tail resembles a straight line in a log-log graph. The head of each graph also resembles a straight line in some cases; however, the head clearly is not part of the same line as the tail, in each case.

The tail of the distribution indicates that we could try to use a power-law distribution or a lognormal to model these graphs. Indeed, we will investigate the feasibility of these two choices in the next section. However, considering the entire graph, we see the need for a distribution that resembles different straight lines in both the head and tail. The DPLN distribution is one such distribution, and we will fit this distribution as well.

5.2 Traditional Fits

5.2.1 Fitting a Power Law

Figure 1: Distribution of number of call-partners per user, in one month, from May 2007

Figure 2: Distribution of per-user calls made in one month, from May 2007

We first attempt to model the observed data using power-law distributions. Like much of the related work[**TODO references**], we attempt to model on the tail of the distribution. Therefore we also have to decide on a truncation point: all data points before the truncation point are not factored into the model. We use the method of maximum likelihoods to obtain the best power-law fit, and associated truncation point. The method we use has been described in detail in prior work[**TODO**] and the code we used was obtained from [**TODO**].

The results of our fitting process for one metric, the number of call partners, are displayed in Figure 4. Clearly, the tail of the data set is well modeled by a straight line in the log-log scale, or a power-law. However, we see that this ignores more than half [**TODO**] the total number of users: all the data points that lie to the left of the truncation point. We obtain similar results, omitted in the interests of brevity, for other metrics and other switches. Therefore we seek to obtain a better model for our data. Indeed, as we will later see, there are models that fit better, not only to the human eye, but also in terms of statistical error metrics.

5.2.2 Fitting a Lognormal

The lognormal distribution has often been considered to be a more descriptive alternative to the power-law distribution [**TODO**]. The shape of the lognormal is a parabola in a log-log scale graph. Since our data plots (from Figures 1, 2 and 3) appear to be closer to horizontal near the head, and steeper near the tail, the lognormal merits investigation as a candidate model.

We obtain the best fitting lognormals using the method of maximum likelihood minimization, as described in [**TODO**], using code from [**TODO**]. The results for one metric, the number of call-partners, are displayed in Figure 5. The lognormal appears to a better fit than the power law, when considering the entirety of data points. However, there clearly is room for improvement. This is substantiated in the next section. Again, the results for other metrics and switches are similar, and omitted due to lack of space.

5.3 Accurate fit

We have observed thus far that our data set exhibits a straight line head and a straight line tail (in a log-log graph of the PDF), and is therefore modeled poorly by the traditionally used power law and lognormal distributions. However, a new distribution, named DPLN, was proposed recently by Reed [**TODO**], which is characterized by a straight line head and a straight line tail (as well as a hyperbolic middle

Figure 3: Distribution of per-user total talk time (in seconds), for one month, from May 2007

Figure 4: DUMMY GRAPH: TODO: replace this by the power law fit graph

Figure 5: Fitting a Lognormal to the distribution of the number of per-user call-partners, for one month, from May 2007

section). This corresponds exactly to what the human eye observes from our data sets! Therefore we now investigate the best fits of DPLN to our data.

Again, we used the method of maximum likelihood, described in [TODO], to obtain the best fits to our data. In some cases, numerical computation and floating point errors limited our ability to find the best fits by this method; in these cases, we obtained a suitable fit by trial and error. As we shall see, even this sub-optimal approach towards fitting DPLN resulted in lower statistical estimation errors than the power law and lognormal fits.

Figure 6 displays our results for one metric, the number of call partners. Clearly, the DPLN is a better fit to our data than the power law and lognormal fits that we observed in Sections 5.2.1 and 5.2.2. This is numerically substantiated in the first column of Table [TODO], which shows the estimation errors of the models we considered.

We performed the same fitting process for other metrics, the number of calls and total call duration, and display these results in Figures 7 and 8. The estimation errors for these metrics are shown in Table [TODO]. Thus, we see that DPLN serves to accurately model our distribution. This result was also true for our data sets collected from other time periods and switches. We illustrate this in the next section.

5.4 Persistence across Time

We now examine call records from a different period of time 6 months later. The records are collected, as before, over a one month period starting from December 2007. Figure ?? shows the histogram of the distribution of call partners for this data set, and the DPLN curve that we fit to this data set. Not only does the DPLN shape persist, but we observe that the parameters of the DPLN distribution are the same as in the May 2007 dataset. This strengthens our argument that the underlying distribution is, in fact, DPLN.

5.5 Persistence across Geographical Areas

Thus far, we have focused on the call records collected from one switch (corresponding to one geographic area: AREA-1). We now examine the data collected from several other areas and time periods, to see if our observations still apply.

Figures [TODO] show the distribution of the per-user number of call-partners observed in a month, collected from

Figure 6: Fitting a DPLN to the distribution of the number of per-user call-partners, for one month, from May 2007

Figure 7: Fitting a DPLN to the distribution of the number of per-user calls, for one month, from May 2007

Figure 8: DUMMY graph: TODO: Replace this by DPLN fit for Seconds

two different switches, corresponding to areas AREA-2 and AREA-3 respectively. The shape of the log-log graph closely resembles its counterpart from AREA-1 (Figure 1), suggesting that the DPLN distribution applies in these cases too. The parameters of the fitted DPLN were obtained as described in Section 5.3, and are shown in Table [TODO]. These parameters are not identical to the parameters obtained in AREA-1; however, there is a high degree of similarity, for example, in the slopes of the graph.

We now investigate the reasons why the DPLN distribution arises in all our datasets.

6. SOCIAL WEALTH: GENERATIVE PROCESS FOR DPLN

In this section, we leverage prior work on social graphs and generative processes to synthesize our model for the generative process underlying our call graphs. We will then use our call data sets to provide evidence supporting our proposed model of a generative process.

Heavy tails have been repeatedly observed in a variety of contexts including graph degrees, filesize distributions, etc. Modeling the entire distribution of these real-world attributes has frequently led to the use of either power law distribution or the lognormal distribution. As pointed out by Mitzenmacher [?], the generative processes that lead to these various distributions are subtly different from each other. In the following survey, we heavily rely on Mitzenmacher [?].

Lognormal distributions are generated by the following multiplicative process. Each individual of a large population has an attribute whose initial value is X_0 . At each step j , the attribute's value increases or decreases based on a random variable F_j :

$$X_j = F_j X_{j-1}.$$

If all the F_j are independent and identical lognormal distributions, then the product is also lognormal and hence, X_j is also lognormal. Even if F_j are independent and identical but not lognormal, the product yields a lognormal distribution asymptotically by the Central Limit Theorem.

Champoernowne's model [TODO-ref] can be viewed as a variation of the above with a minimum value for the attribute X_j . This seemingly-minor variation provides a power law distribution instead of a lognormal distribution.

Often, while considering income distributions, etc. we ob-

Figure 9: Fitting a DPLN to the distribution of the number of per-user call-partners, for one month, from Dec 2007

Figure 10: Fitting a DPLN to the distribution of the number of per-user call-partners, for one month, from Dec 2007, for a different Area - AREA2

Figure 11: Fitting a DPLN to the distribution of the number of per-user call-partners, for one month, from Dec 2007, for a different Area - AREA3

serve the attributes of different individual at different times of their lives. This leads to the examination of X_T where T itself is a random variable. In the case where T is an exponentially distributed random variable, Reed [?] shows that the resulting distribution is a double Pareto distribution with two Pareto tails. Reed’s dPIN model is a generalization of the double Pareto distribution that further assumes that the initial value X_0 is not the same but is also lognormally distributed.

This distribution was found to model income distribution accurately, and a corresponding generative process was proposed in [?]. Reed [?] explained that three major assumptions were required for a generative model resulting in a DPLN distribution of incomes: (a) Starting incomes are lognormally distributed; (b) People’s earning lifetimes are exponentially distributed; (c) Individuals’ income growth is governed by geometric brownian motion [TODO-ref].

We hypothesize that social graph attributes are a measure of *social wealth* and can be modeled by dPIN in much the same way as actual incomes. Of the three assumptions listed above, we suggest that social wealth satisfies (a) and (b) for the same reasons that Reed does; we do not provide evidence specific to phone calls, for this purpose. However, we provide two key empirical observations to justify our hypothesis:

1. *Metrics of social wealth including degrees, number of phone calls and number of minutes talked all can be well modeled by the dPIN distribution, with persistent parameters across different time periods (demonstrated in Section 5).*

and

2. *The evolution of the above metrics across a time period appears to be well modeled by a lognormal multiplicative process (thus satisfying assumption (c) above).*

To analyse the process of evolution of social wealth (i.e. Observation 2 above), we examine the the ratio of the number of call-partners for the same set of users in two different months: Dec 2007 and May 2007. The histogram of the distribution of *this ratio* is plotted in Figure 12 in the log-log scale, and appears to have a parabolic shape, corresponding to the lognormal distribution. Indeed, the lognormal distribution was found to provide a good fit, as shown by the solid line in Figure 12. Similarly, the ratio of number of calls and call duration for two datasets 6 months apart appeared to be distributed lognormally, as shown in Figures 13 and 14 respectively.

Figure 12: Distribution of Ratio of Node Degrees in Dec’07 to May ’07

Figure 13: Distribution of Ratio of Per-user Calls in Dec’07 to May ’07

Figure 14: Distribution of Ratio of Per-user total call-duration for a month, in Dec’07 to May ’07

[TODO] Table - Observations of all the remaining ratios 6 months apart.

7. CONCLUSIONS AND FUTURE WORK

0.5 page

Power law distributions and the processes that generate them are widely believed to characterize many real-world phenomena. In this paper, we analyzed user behavior in a large social network at a mobile phone operator, consisting of millions of users and tens of millions of calls over different time periods, and found evidence suggesting fundamentally different characteristics. In particular, we found the following:

- Though per-user characteristics such as degree and number of calls have power law tails, their distributions were not well-modeled by power law or lognormal distributions in their *entirety*.
- With few exceptions, all the characteristics in our call graph were well-modeled by a recently formulated heavy-tailed distribution, the *double Pareto log Normal (dPIN)* distribution, which is characterized by two linear log-log components.
- Over time, our graph showed an evolution that was consistent with a generative process that is based on geometric Brownian motion and leads to dPIN distributions. Furthermore, this generative process, which has often been used to explain the dPIN nature of income distributions, appears to lend itself to a natural and appealing *social wealth* interpretation in the context of social networks such as ours.

Our results question the applicability of traditional power law based processes and models in explaining social networks and demonstrate the potential superiority of dPIN in the same context. Though our work is not broad enough to comprehensively answer these questions, it is intended to spur more studies into these questions. In particular, we hope that our “social wealth” hypothesis and analysis serves as incentive for social scientists to study the large-scale evolutionary aspects of social characteristics. Indeed, we ourselves continue to collect data from our social network for longer-term analysis.

bib 0.5 page