Mobile Call Graphs: Beyond Power-Law and Lognormal Distributions [*]

Mukund Seshadri
Sprint
Burlingame, California, USA
mukund.seshadri@sprint.com

Sridhar Machiraju
Sprint
Burlingame, California, USA
machiraju@sprint.com

Ashwin Sridharan
Sprint
Burlingame, California, USA
ashwin.sridharan@sprint.com

Jean Bolot
Sprint
Burlingame, California, USA
bolot@sprint.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 the records of a large mobile phone operator, with more than a million users and tens of millions of calls. We examine the distributions of the number of phone calls per customer; the total talk minutes per customer; and the distinct number of calling partners per customer. We find that these distributions are skewed, and that they significantly deviate from what would be expected by power-law and lognormal distributions.

To analyze our observed distributions (of number of calls, distinct call partners, and total talk time), we propose *PowerTrack* , a method which fits a lesser known but more suitable distribution, namely the Double Pareto LogNormal (DPLN) distribution, to our data and track its parameters over time. Using *PowerTrack* , we find that our graph changes over time in a way consistent with a generative process that naturally results in the DPLN distributions we observe. Furthermore, we show that this generative process lends itself to a natural and appealing *social wealth* interpretation in the context of social networks such as ours. We discuss the application of those results to our model and to forecasting.

Categories and Subject Descriptors

Database Management [**Database Applications**]: Data Mining

General Terms

Algorithms, Experimentation

Keywords

Distribution, Generative Process, Power Laws, DPLN

1. INTRODUCTION

Conventional wisdom now holds that power-law distributions and the processes that generate them are ubiquitous in an extremely wide range of phenomena, ranging from "real-world" or physical phenomena and constructs, such as the degrees of proteins or the number of species per genus of mammals, to "virtual" phenomena and constructs such as the degree of nodes in the Internet or the number of citations received by papers (e.g., see [8] or [18] for an analysis of two dozen data sets from a wide variety of fields). The results that led to this state include an extremely large collection of results showing that i) measured data of the above phenomena do exhibit heavy-tailed distributions, especially power-law and lognormal distributions, and ii) simple generative processes such as preferential attachment can be used to understand and explain the reasons for the ubiquity of heavy-tailed distributions in the natural and the virtual worlds [15, 18].

We focus in this paper on the analysis of the social network formed by the calls of users in phone networks. The behavior of users in landline networks has been examined, for example, by considering communities of interest among those users [9]. Of particular interest to us, in this paper, is the analysis of the social network formed by the phone calls of mobile users in cellular networks. The analysis of mobile phone graphs is an exciting area of research, because mobile phones are ubiquitous, they have become a strategic component of modern life and modern economies, and they are expected to become a key or even the principal conduit not just for voice calls, but for Internet access and use in the future as well [13]. Furthermore, they can provide detailed information on the spatio-temporal behavior of users, especially on their mobility patterns and on the social networks they build and maintain, as reflected by their phone calls. Several recent studies have used mobile call graph data to examine and characterize the social interactions of cell phone users, with a focus on understanding the structural properties of the graph [12, 17], its evolution and the

[*] This material is based upon work partially supported by the National Science Foundation under Grant No. IIS-0534205, and a Sprint gift. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, or Sprint.

KDD'08, August 24–27, 2008, Las Vegas, Nevada, USA.
Copyright 2008 ACM 978-1-60558-193-4/08/08 ...$5.00.

evolution of social groups [19], or the spread of new products and services [25].

In this paper, we examine the mobile call graph, and the corresponding social networks, obtained from the network of a large cellular operator. The network we consider is a geographical subset of a continent-wide network with several dozens of millions of users and several billions of calls, where even the subset involves a million users and tens of millions of calls. We examine the distributions of the number of phone calls per customer; the total talk time per customer; and the distinct number of calling partners per customer. We also observe how those distributions might differ at different points in time. A relatively small number of studies of similarly-scaled networks have been reported in the literature (in particular [17, 12]) and, consistent with the conventional wisdom mentioned above, have reported power-law distributions for measures such as degree distribution, etc.

Our contributions are fourfold. First, we find that the distributions in our dataset significantly deviate from those observed in earlier work, and that traditional power laws often fall short. Second, we introduce our *PowerTrack* method which provides significantly better fits using the lesser known but more suitable *Double Pareto LogNormal (DPLN)* distribution. *PowerTrack* neatly summarizes an observed data distribution using four parameters, which we can easily compute at any given point in time, and monitor over time. Third, we find that our graph changes over time in a way consistent with a generative process that naturally results in the DPLN distributions we observe. And fourth, we show that this generative process lends itself to a natural and appealing *social wealth* interpretation in the context of social networks such as ours.

The rest of the paper is organized as follows. In Section 2, we provide background information on heavy-tailed distributions and review related work relevant to the paper. In Section 3, we describe the dataset used in the paper, develop the *PowerTrack* methodology, and present the results of analyzing our dataset using typical heavy-tailed distributions, namely, the power-law and lognormal distributions. In Section 4, we apply *PowerTrack* to our dataset, provide evidence for a fit with the DPLN distribution, examine how the parameters of the distribution evolve over time, and discuss several practical applications of our results. In Section 5, we derive a generative process based on social wealth to explain the experimental results of Section 4, and we discuss the implication of our findings for network and social scientists. Section 6 concludes the paper.

2. BACKGROUND

In this section, we provide background on skewed distributions, and in particular, power-law and lognormal distributions. We also survey prior work in these areas.

Power Laws

Power laws, which have been observed in an overwhelming number of settings including graphs and social networks, are characterized by the following probability distribution:

$$f(x) = Cx^{-\alpha}\,, \tag{1}$$

Examples of power-law degree distributions in graphs include the Internet AS (Autonomous System) graph with exponent $\alpha = 2.1 - 2.2$ [10], the Internet router graph with exponent ~ 2.48 [10, 11], the in-degree and out-degree distributions of subsets of the world wide web with exponents 2.1 and $2.38 - 2.72$ respectively [3, 14, 6], the in-degree distribution of the African web graph with exponent 1.92 [5], a citation graph with exponent 3 [22], distributions of website sizes and traffic [1], and many others. Newman [18] provides a comprehensive list of such work.

Deviations

While power laws appear in a large number of graphs, deviations from a pure power law are sometimes observed. Pennock et al. [21] and others have observed deviations from a pure power-law distribution in several datasets. Two of the more common deviations are exponential cutoffs and lognormals. In exponential cutoffs, the distribution looks like a power law over the lower range of values along the x-axis, but decays very fast (exponentially) for higher values. Amaral et al. [2] find such behaviors in the electric power-grid graph of Southern California and the network of airports, the vertices being airports and the links being non-stop connections between them.

Lognormals or the "DGX" distribution

The lognormal distribution is a parabola in log-log scales, but may seem like a power law, if appropriately masked. Pennock et al. [21] recently found while the whole WWW does exhibit power-law degree distributions, subsets of the WWW (such as university homepages and newspaper homepages) deviate significantly. They observed unimodal distributions on the log-log scale. Similar distributions were studied by Bi et al. [4], who found that a discrete truncated lognormal (called the Discrete Gaussian Exponential or "DGX" by the authors) gives a very good fit. A lognormal is a distribution whose logarithm is a Gaussian. The DGX distribution extends the lognormal to discrete distributions (which is what we get in degree distributions), and can be expressed by the formula:

$$f(x = k) = \frac{A(\mu,\sigma)}{k} \exp\left[-\frac{(\ln k - \mu)^2}{2\sigma^2}\right] \quad k = 1, 2, \ldots \tag{2}$$

where μ and σ are parameters and $A(\mu, \sigma)$ is a constant (used for normalization if $f(x)$ is a probability distribution). The DGX distribution has been used to fit the degree distribution of a bipartite "clickstream" graph linking websites and users, telecommunications and other data.

Mobile Call Graphs

In the next section, we show that none of these variations provide a best-fit to our real-world dataset of mobile phone calls, thus motivating our core objectives in developing *PowerTrack*. Mobile networks have been previously analyzed in literature, *e.g.,* Onnela *et. al.* [12], Nanavati *et. al.* [17]. However, their focus was on characterizing an aggregate snapshot of the network in terms of neighbor distribution, topology, social interaction, *etc.* We differ in that we not only analyze user behavior, but also postulate an underlying causal temporal generative process and test it against temporally diverse datasets. Our results demonstrate a calling behavior that has not been previously analyzed in the context of call graphs. Furthermore, we leverage the temporal aspects of our data to gain insight into salient features of the user calling process.

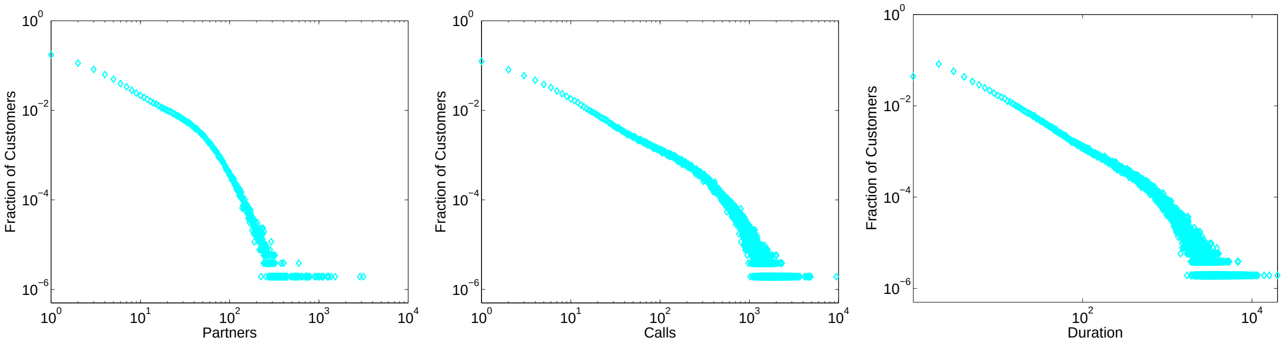

Figure 1: Distribution in time period $T1$ of *Partners* (Left), *Calls* (Middle) and *Duration* in minutes (Right), for users at S1.

3. GOALS AND PRELIMINARY OBSERVATIONS

We start this section by describing the dataset used in this paper. We then present the goals and the design rationale behind our proposed *PowerTrack* methodology. Finally, we present the result of analyzing our dataset using the heavy-tailed distributions typically used in the past to analyze large graphs, namely the power-law and the lognormal distributions.

3.1 Dataset

The dataset analyzed in this paper is made of a large collection of Call Data Records (CDRs) from a large cellular network. This network supports voice, data, and SMS services, and the CDRs include information about all 3 types of services. However, we only consider voice calls in the paper.

CDRs were collected at several Base Station Controllers (BSCs). The collection function is provided by the equipment deployed in the network, as part of the normal troubleshooting and billing capabilities. We collected call records at four different switches, which we refer to as $S1$, $S2$, $S3$ and $S4$. Each of these switches recorded calls made to and from callers who were physically present in a contiguous geographical area. The areas covered by the four switches were also geographically contiguous. Apart from geographical diversity, we also incorporated temporal diversity by collecting records at switch $S1$ during two (month-long) time periods, $T1$ and $T2$, which were separated by 6 months.

Each of our month-long datasets at any single switch collected roughly $20-50$ million call records from and to about half a million [1] mobile users when they were within its geographic area. Throughout this paper, we only focus on calls that terminated successfully.

Call Data Records include several fields about each call event. Of interest in this paper are the calling and called parties (source and destination of calls), and the duration of calls. We emphasize that our interest is in aggregate statistical analysis and therefore, we do not study any particular individual's calling pattern. More importantly, in order to maintain privacy and anonymity, data that could identify users (e.g. their phone numbers) is not utilized in this study; we analyze anonymized CDRs, and restrict our focus to the patterns of calls and networks formed out of these calls.

3.2 Goals and Design of *PowerTrack*

From the data described above, we obtain a call graph G which is a tuple (V, E) where V denotes a set of vertices, representing the mobile users, and E denotes a set of edges, representing the mobile calls. Specifically, if x and y are vertices of G, then an edge exists between x and y if x and y have called each other at least once during the time interval of interest. We represent multiple calls between any two nodes by a single edge, which can be associated with a weight (equal to one to represent connectivity, or equal to the total number of calls or the number of minutes between the two nodes during the interval of observation). In this paper, we assume undirected edges, i.e. we do not distinguish between callers and callees.

Our goal, then, is to analyze our graph G, and specifically to characterize the underlying behavior of mobile users and derive insight into how and why the observed characteristics arise. To achieve our objective, we utilize a three-step methodology that we refer to as *PowerTrack*.

The first step involves choosing three instructive per-user characteristics that measure the behavior of individual users in the underlying social graph, namely,

- ***Partners*:** The total number of *unique* callers and callees associated with every user. Note that this is essentially the degree of nodes in the (undirected and unweighted) social graph, which has an edge between two users if either called the other. In addition, we use the term "call partners" to refer to the set of unique callers and callees associated with a user.

- ***Calls*:** The total number of calls made or received by each user. In graph theoretic terms, this is the weighted degree in the social graph where the weight of an edge between two users is equal to the number of calls that involved them both.

- ***Duration*:** The total duration of calls for each customer in minutes. This is the weighted degree in the social graph where the weight of the edge between two users is the total duration of the calls between them.

Each of the above metrics can be calculated, per user, from the call records over any period of time. Prior studies [17] have used a month of calling behavior to characterize user characteristics. Furthermore, many phone calling plans are based on monthly usage. Hence, we decided to use a period

[1] We only provide approximate numbers for proprietary reasons.

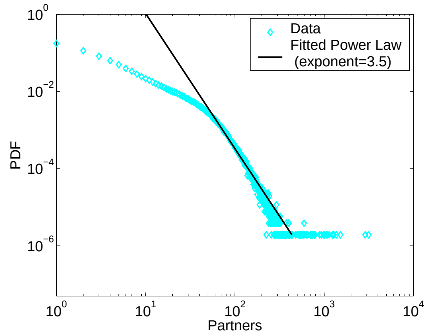

Figure 2: Power-law fit to the distribution of *Partners* per user, for $S1$ during $T1$.

of one month to calculate the per-user metrics. We also decided to analyze these metrics by studying their probability distribution functions (PDFs).

In the second step (detailed in Section 3.3 and Section 4), we analyze the observed distributions of our chosen metrics and derive statistical distributions that best fit the empirical distributions. We explore the nature of the best-fit distributions across geographically diverse datasets as well as temporal instances to gain a better understanding of the social graph.

In the third step (detailed in Section 5), we use our data collection over time to gain insights into the generative processes that best describe the temporal evolution of the user calling behavior.

3.3 Power-law and Lognormal Fits

As discussed above, the second-phase of *PowerTrack* involves using our datasets to estimate the empirical density functions and fitting them to the standard statistical distributions we expect (based on the shape of the distribution curves and on past work on graph data analysis) will work best. For each metric X, we choose a bin size b and estimate the probability distribution function (PDF) of the metric X at $x = b, 2b, 3b \cdots$ as:

$$\hat{p}(x) = \frac{\|X \in [x - \frac{b}{2}, x + \frac{b}{2}\|}{\|X\| \cdot b}. \tag{3}$$

where the R.H.S. uses the empirical probability of observing X within the interval. The bin size is chosen to be large enough so that this empirical probability can be well estimated. In Figure 1, we plot the densities of our three metrics estimated using the dataset from switch $S1$ during time period $T1$. All three figures are plotted in the log-log scale. Not surprisingly, we notice that all densities have a heavy tail which is clearly linear in the log-log scale. The heavy tail provides motivation to model the densities above using two well-known distributions, namely the power-law and lognormal distributions.

We first attempt to model the observed data using power-law distributions (see Equation 1). Following the lead of prior work [8], and using code from [7], we determine a power-law fit by using Maximum Likelihood Estimation (MLE) to model the tail of the distribution. This power-law fit consists of two parameters - a truncation point that defines the tail and the exponent of the power-law fit to the tail. The power-law distribution that best models the distribution of the *Partners* metric is shown in Figure 2.

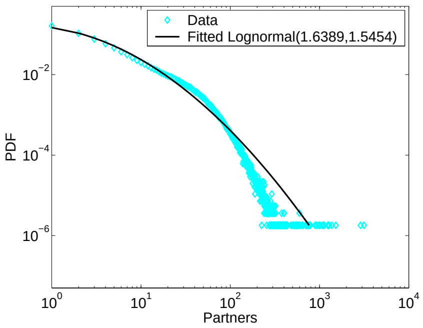

Figure 3: Lognormal fit to the distribution of *Partners* per user, for $S1$ during $T1$.

Note that, for our metrics of interest, we cannot observe any data point with values less than one. Hence, we truncate our empirically observed distributions at one. Here and later in this paper, to fit an analytical model that is (in some cases) supported from zero to infinity, we scale our PDF estimate in Equation 3 so that the areas under the empirical and modeled PDF curve are the same.

While the power-law best-fit in Figure 2 models the tail of the distribution well, it does a poor job of modeling the head of the distribution, which is to the left of the truncation point. We obtain similar results when we try to use power laws to fit the distributions of the other two metrics.

Given the inability of power laws to fit the head of our distributions, we then turn to lognormal distributions (Equation. 2), which have often been seen as a good alternative to power laws. Like the power-law distribution, the lognormal distribution also has an almost linear tail. However, lognormal distributions have a parabolic shape in the log-log scale, which appears to be similar to the shape of the distributions in Figure 1. We obtain the best fitting lognormals using Maximum Likelihood Estimation, as described in [4]. The best lognormal fit to the distribution of *Partners* is shown in Figure 3. Though this appears to a better fit than the power-law fit, it can clearly be improved. We achieve similar (negative) results trying to fit power-law and lognormal distributions with the other metrics using this dataset, and with the other datasets as well.

4. ANALYSIS OF THE DISTRIBUTIONS

In the previous section, we discussed the results of using well-known heavy-tailed distributions - power-law and lognormal - to fit the distributions of our chosen metrics. We found that the best-fits show clear scope for improvement especially in modeling the head of the distribution. In this section, we present the first result from *PowerTrack*, namely, that a recently-formulated Double Pareto Log Normal (DPLN) distribution yields good fits to our empirical distributions. We start by providing a quick introduction to the DPLN distribution and discuss its salient properties to motivate its selection as a best-fit for our empirical distributions. Most of the discussion is based on the work done by Reed [23].

4.1 The DPLN Distribution

The DPLN distribution arises out of a mixture of lognormal distributions as described below. Consider a random

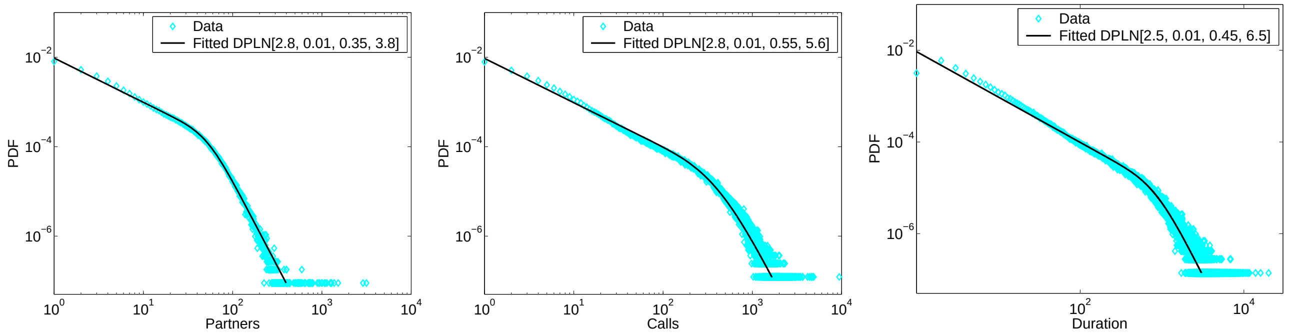

Figure 4: Results of using DPLN to model *Partners* (Left), *Calls* (Middle) and *Duration* in minutes (Right), for users at S1 in the time period $T1$.

variable representing the state S of a Geometric Brownian Motion (GBM), i.e., the logarithm of S follows Brownian motion and, hence, satisfies:

$$dS_t = \mu S_t dt + \sigma S_t dw_t \tag{4}$$

Here, w is the Weiner process and μ, σ are constants. The initial state S_0 is considered to be lognormally distributed with an underlying normal distribution $N(\nu, \tau^2)$. After T time units, the state S is also distributed lognormally with an underlying normal distribution:

$$S_T \sim LN(\nu + (\mu - \frac{\sigma^2}{2})T, \tau^2 + \sigma^2 T). \tag{5}$$

Moreover, $\frac{S_T}{S_0}$ is also lognormal:

$$\frac{S_T}{S_0} \sim LN(\mu - \frac{\sigma^2}{T}, \frac{\sigma^2}{T}). \tag{6}$$

If the observation time T is exponentially distributed with parameter λ, then the random variable $X = S(T)$ has a DPLN distribution denoted as $DPLN(\alpha, \beta, \nu, \tau)$ where ν and τ are as above and $\alpha > 0$ and $-\beta < 0$ are roots of the quadratic equation:

$$\frac{\sigma^2}{2} z^2 + (\mu - \frac{\sigma^2}{2})z - \lambda = 0\,. \tag{7}$$

The complete DPLN distribution is given by:

$$\begin{aligned} f(x) \quad &= \tfrac{\alpha\beta}{\alpha+\beta}\Big[e^{(\alpha\nu+\alpha^2\tau^2/2)}x^{-\alpha-1}\Phi(\tfrac{\log x - \nu - \alpha\tau^2}{\tau}) + \\ &\quad x^{\beta-1}e^{(-\beta\tau+\beta^2\tau^2/2)}\Phi^c(\tfrac{\log x - \nu + \beta\tau^2}{\tau})\Big]\,, \end{aligned} \tag{8}$$

where Φ and Φ^c are the CDF and complementary CDF of $N(0, 1)$.

An easier way of understanding the double Pareto nature of X is by observing that $X = S_0 \frac{V_1}{V_2}$ where S_0 is lognormally distributed, and V_1 and V_2 are Pareto distributions with parameters α and β. Note that X has a mean that is finite only if $\alpha > 1$ in which case the mean is given by

$$\frac{\alpha\beta}{(\alpha-1)(\beta+1)} e^{\nu + \frac{\tau^2}{2}}.$$

The distinguishing features of the DPLN distribution are two linear sub-plots in the log-log scale and a hyperbolic middle section. These bear a striking similarity to our empirical distributions in Figure 1 thereby motivating our exploration of DPLN to model them. We explore this in greater detail in the next sub-section(s) and the role of the temporal generative process (Equation 4) in Section 5.

4.2 DPLN Best Fits

In order to estimate the parameters $(\alpha, \beta, \nu, \tau)$ of the DPLN distribution that best fits our empirical data, we initially explore the method of Maximum Likelihood Estimation, described in [23]. We find that this method is sensitive to numerical computation issues, especially floating point rounding off errors. Hence, in some cases, we manually obtain a DPLN fit to our empirical distributions, by performing a grid search of the parameter space. Though such a manual approach is not scalable, it suffices for our purpose of illustrating its superiority over other distributions. Obtaining a practical automated fitting method is an important area of future work.

In Figure 4, we plot our DPLN best fits for the plots in Figure 1. We plot the best fits based on code from the authors of [23]. The DPLN is clearly seen to be a better fit to our data than the power-law and lognormal fits that we derived earlier. We also numerically substantiate this by quantifying the difference between the analytical and empirical distributions via the Residual Sum of Squares (RSS) using geometric binning. The RSS value for DPLN is 9.8×10^{-6}, which is two orders of magnitude smaller than the RSS of 2.9×10^{-4} for the best lognormal fit. Like us, Mitzenmacher [16] also obtained good fits using DPLN, for file size distributions.

4.3 Temporal Diversity

We now examine call records for $S1$ from our second month-long time period $T2$, which was 6 months after $T1$. Figure 5 (**Left**) shows the empirical and DPLN best-fit density function for *Partners*, for this data set. Not only does DPLN continue to model the empirical dataset well, the parameters of the best-fit DPLN distribution do not change significantly from $T1$ to $T2$. Since users may move out of or leave the network during the 6-month period between $T1$ and $T2$, we observe a different set of users during $T2$. The persistence of the best-fit DPLN parameters, in spite of a dynamic set of users, suggests that the DPLN distribution might arise due to fundamental large scale social network characteristics. We will explore this further in Section 5.

4.4 Spatial Diversity

Thus far, we have presented results that used only the call records collected from $S1$. To verify that the DPLN na-

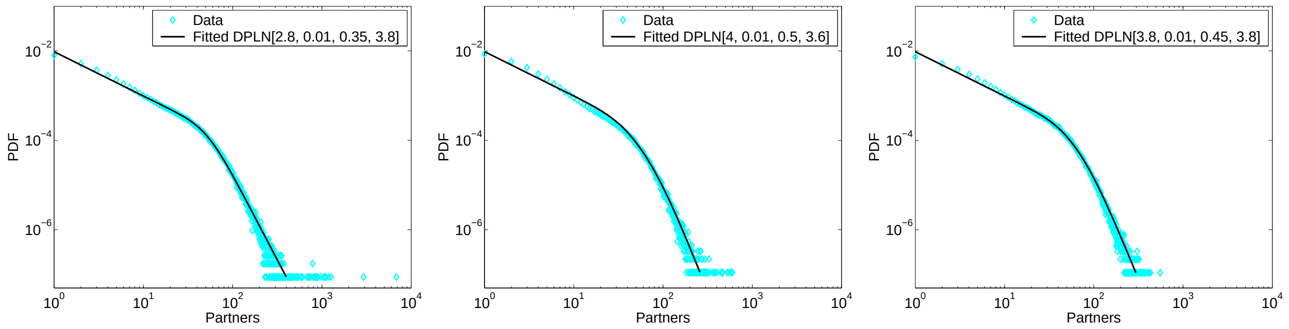

Figure 5: Results of using DPLN to model *Partners* during $T2$ for area $S1$ (Left), $S2$ (Middle) and $S3$ (Right).

ture of the empirical distribution was not specific to just a single area, we examined the data collected from the three other areas. Figure 5 (**Middle** and **Right**) shows the DPLN best-fits for two such areas during the time period $T2$. These plots confirm the general applicability of DPLN to our datasets. However, the actual best-fit parameters do vary across the different areas as do the empirical distributions. The demographics and living standards in the area covered by $S1$ are quite different from that of $S2$ and $S3$. This is good empirical evidence that the nature of social graphs is not uniform and can vary significantly, though similar statistical laws may still apply.

4.5 Other Applications

We now briefly describe ways in which to take advantage of the DPLN fit to our observed data. While several applications are possible, we focus on two in particular: an application to outlier detection, and an application to workload management.

4.5.1 Outlier Detection

So far, we have measured the per-user call durations in units of minutes. Given our success with DPLN fits using *PowerTrack*, we are motivated to investigate the distribution of per-user total call durations in units of *seconds*, too. In Figure 6, we plot this distribution. We find that it is also well-described by a DPLN model (shown in Figure 6). However, we find two outliers at 27 and 54 seconds. *PowerTrack*'s ability to fit the rest of the distribution implies that these are genuine outliers worthy of investigation. Indeed, we found that these outliers arose due to a common exceptional scenario in the calling process, namely, when a mobile user did not answer an incoming call and the caller hung up without leaving a voicemail. The 54 seconds represent users who received exactly 2 such calls during the time of observation. While this is a specific scenario, it serves to illustrate the applicability of our model to outlier detection in general.

4.5.2 Pricing Structure Design

Our results so far provide us with an accurate model of the workload generated by mobile phone users at large time scales. In particular, we have a model of the distribution of user behavior (in terms of total duration of mobile phone use) over a month. Such models can be used to design pricing structures that charge users differently according to their "tiers" of monthly usage (such structures are common today). For example, our model can maximize total revenue by helping us determine the amount of money each user (or group of users) is charged. This could be traded off against the cost of supporting users, to optimize measures of marginal gain. Furthermore, our models could be used as input to solving a dynamic system. The pricing structure impacts the rate at which customers sign up or leave different billing plans. Therefore it would have to account for the resultant dynamic workload, with the aim of fueling as much growth in customer base (and revenue) as underlying network resources can support. Queueing models to address such systems are common and could utilize our models of user workload to determine optimized billing decisions.

5. SOCIAL WEALTH: GENERATIVE PROCESS FOR DPLN

In the previous section, we showed the results obtained using the second step of *PowerTrack*, namely, the derivation of best-fits for our observed data. We found that *PowerTrack* yielded valuable information about the user characteristics in our social graph as well as its temporal and spatial variations. In this section, we present the results of *PowerTrack*'s third step - using data collected over different time periods to understand the underlying generative process of our social graph. We start by surveying prior work in power-law and lognormal generative processes. Then, we describe our social wealth based generative process and provide substantial evidence supporting it.

5.1 Proportional Effects

A great deal of work (see [15] for exhaustive references) has been done to understand how heavy-tailed distributions such as power-law and lognormal arise. Gibrat [24], for example, proposed the law of proportional effects to understand the distribution of sizes of industrial firms. The core principle behind this law is that the growth of a firm is multiplicative and independent of its current size. In other words, if X_j represents the size of a firm at a discrete time step j,

$$X_j = F_j X_{j-1} \tag{9}$$

where F_j is a random variable independent of X_{j-1}. If the F_j are independent and identically distributed random variables, the Central Limit Theorem can be used to show that X_j is asymptotically lognormal. In fact, if the F_js are themselves lognormal, then X_j is always lognormal. This multiplicative model and variants based on it have been used to lognormally model a wide variety of real-world attributes (see [15] for a survey).

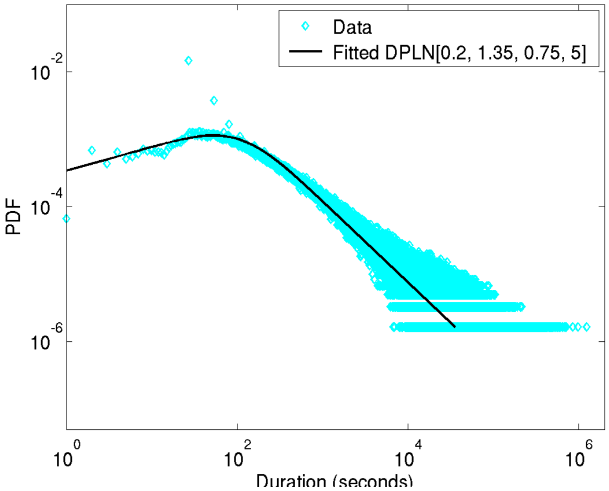

Figure 6: Results of using DPLN to model the total call Duration measured in *seconds* during $T1$ for area $S1$.

As noted in [15], a simple modification to the model above leads to power-law distributions. Specifically, if the X_j values are lower bounded by a minimum value, then the resulting distribution turns out to be a power-law instead of a lognormal. Indeed, Pareto [20] introduced the Pareto distribution by a similar process to explain tail income distribution. Note that power laws have been used to model attributes such as node degrees in many real-world graphs. Such modeling has been done by describing the graph evolution using a process of *preferential attachment* [3], which is similar to proportional effects. Preferential attachment states that new nodes attach themselves to existing nodes with a probability that is proportional to their degrees.

5.2 Social Wealth

We start our investigation into the generative process underlying our social graph using data from time periods $T1$ and $T2$. Some of the users observed during $T1$ are not observed in $T2$ and vice versa, either because they physically moved away or are no longer subscribers. We eliminate all such users from our analysis. For the remaining users, in the spirit of Equation 9, we calculate the ratio $\frac{X_{T1}}{X_{T2}}$ for each of our three metrics, where X_t is the user metric during time period t. We plot the distribution of these ratios in Figure 7.

As seen in Figure 7, the distributions of the ratio appear to be parabolic on the log-log scale. Hence, we use Maximum Likelihood Estimation to fit these distributions to lognormals. We find remarkably good fits for all the distributions. The parameters of these best-fits are also shown in Figure 7. These results provide good evidence that a *lognormal multiplicative process* is behind the temporal evolution of our social graph.

As discussed in Section 5.1, lognormal multiplicative processes (see Equation 9) have been successfully used to model income distributions. Using these as motivation, we hypothesize that our metrics (*Partners*, *Calls* and *Duration*) capture *social wealth*, the social analogue of income. We believe that our social wealth interpretation provides a natural and appealing extension of income to the social context, and can potentially be used to better understand social behavior in many contexts (e.g., phone networks, the Internet, email networks). We offer two key arguments to support our social wealth interpretation.

Our first argument is consistency with Gibrat's law of proportional effects. The results in Figure 7 not only provide evidence that there is a lognormal-based multiplicative process but also show that such a process accurately models the generative process of social wealth as captured by *any* of our metrics. Another important aspect of the law of proportional effects is the independence between the multiplicative factor (F_j in Equation 9) and the current attribute (X_{j-1} in Equation 9). For our social wealth interpretation, this is equivalent to independence between *Partners* (or *Calls* or *Duration*) for users during $T1$ and the ratio of *Partners* (or *Calls* or *Duration*) across two time periods $T1$ and $T2$. Since the demonstration of independence is difficult, we use cross-correlation (which is a necessary but not sufficient condition for independence). We find cross-correlation coefficients to be uniformly small: -0.14, -0.06 and -0.02 for *Partners*, *Calls* and *Duration* respectively.

As discussed in Section 5.1, lognormal multiplicative processes result in lognormal distributions. Recall though, from Section 4.1, that DPLN distributions arise when a random variable, which has a lognormally-distributed initial value and evolves according to a lognormal multiplicative process, is observed at exponentially distributed random observation times. In other words, if X_t evolves according to a lognormal multiplicative process and T is exponentially distributed, then X_T is DPLN. Thus, in the framework of our social wealth interpretation, our data reflects the social wealth of users who are at different stages in their lifetime, which is assumed to be exponentially distributed. With this interpretation, the best-fit DPLN distributions achieved using *PowerTrack* would be consistent with the notion of social wealth.

While Figure 7 is consistent with the lognormal multiplicative process of DPLN, it is not possible for us to verify if initial values of social wealth are indeed lognormally distributed. In fact, since babies rarely use mobile phones on their own, it may be impossible to directly capture their social wealth. However, the DPLN fits provided us with a way to estimate user lifetimes, which in turn can be used to judge the legitimacy of the best-fit. Consider Equation 7. Since its roots are α and $-\beta$, we have:

$$\alpha\beta = \frac{\lambda}{\frac{\sigma^2}{2}} \tag{10}$$

$$\frac{1}{\lambda} = \frac{2}{\alpha\beta\sigma^2} \tag{11}$$

PowerTrack's best-fits provide us with estimates of α and β. The variance of the fitted lognormal distributions in Figure 7 is $\frac{\sigma^2}{T}$ from Equation 6 where T is 6 months. Thus, we can estimate the average lifetime of users under our generative process. We obtain lifetime values of 43, 23 and 15 years for *Partners*, *Calls* and *Duration* respectively. These values are close to actual average lifetimes, up to a small multiplicative factor of 2 or 3. This inaccuracy may be due to the small values of β, which make estimates of lifetimes highly variable, or it might indicate that social wealth accumulation (in terms of our metrics) starts only at adulthood. Still, the lifetime values we obtain are surprisingly close to actual human lifetimes and offer substantial supporting evidence for our social wealth interpretation.

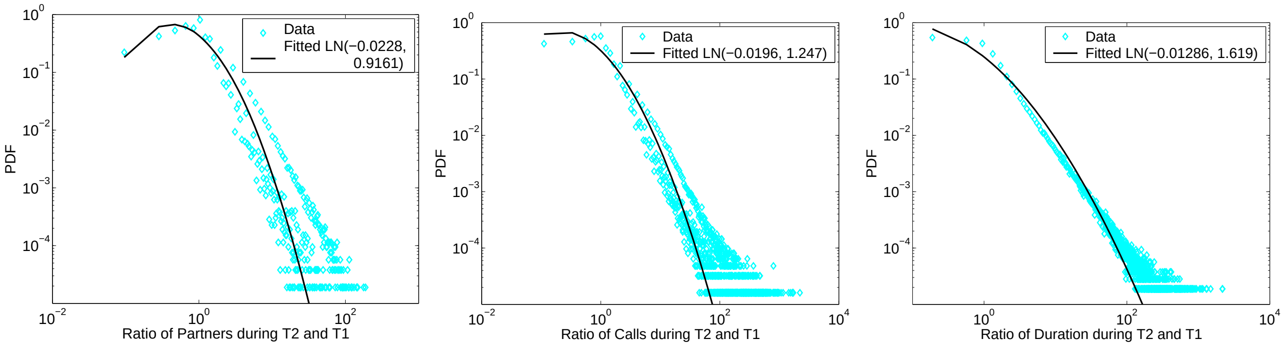

Figure 7: Demonstration of the consistency with our social-wealth generative process. The ratio of user degrees (Left), number of calls (Middle), and, total talked time (Right), in month $T2$ and month $T1$.

5.3 User Calling Patterns

In the previous sections, we explored ensemble call behavior patterns. In this section, we take a closer look at the *type* of users that constitute the population. Specifically, we categorize users according to their social behavior across time, by analyzing the persistence of the set of call partners of each user.

For a given user u, let $S(T_1)$ be the set of call partners during time-period T_1 and $S(T_2)$ the set of call partners during time period T_2. To track the evolution of the social network of a user, we define a metric, *New Caller Ratio* as :

$$\eta \;=\; \frac{|S(T_1) \cup S(T_2)|}{|S(T_1)| + |S(T_2)|}\;. \qquad (12)$$

The above metric is always less than one and captures a key aspect of the evolution of users' call partners. If a user calls disjoint sets of users in $T1$ and $T2$, then $\eta = 1$ (this also occurs when the user is active in only one of the two time instances). However, if a user has a persistent set of call partners, then the metric would be less than one, decreasing with the number of persistent call partners.

Figure 8 (**Top**) shows the distribution of η computed across time instances $T1$ and $T2$, for *all* users at switch $S1$. Figure 8 (**Bottom**) shows the same quantity, for only those users that were active during both $T1$ and $T2$. It is also instructive to observe the distribution of the call partner set sizes (i.e., the *Partners* metric) for these users, labeled "Data PDF" in both plots.

The plots exhibit an interesting, and somewhat unexpected trend. Specifically, the upper right section of both plots indicate that the highest values of η are exhibited by customers with the largest *Partner* values. Intuitively, a residential customer would not typically call a large number of new contacts every month. Such calling patterns are more typical of "robots", e.g. telemarketers, and spammers; verification of this hypothesis is an area of future work.

Figure 8 (**Top**) and (**Bottom**) differ primarily in the lower ranges of the X axis. In combination with the distribution of *Partners*, this indicates the existence of a large number of short-lived users who are associated with small but extremely dynamic sets of call partners. Furthermore, the trends in the distribution of η and *Partners*, in Figure 8 (**Top**), indicate that the left and right tails of the *Partners* distribution might be largely comprised of dynamic short-lived users and the aforementioned "robots", respectively.

In summary, by tracking the values of η and *Partners*, we can potentially distinguish between several important classes of customers: atypical "robotic" customers like telemarketers, characterized by large *Partner* and η values; and typical residential customers, who comprise the remainder. The latter set can be further categorized by identifying the subset of dynamic short-lived customers. These observations also have applications to user clustering in the social graph context, since we expect the residential class of users to form cliques or well-connected clusters, while the atypical users with large *Partner* values would appear as hubs of large-degree "spokes". Further study of these applications, as well as the impact on our observed distributions, requires collecting more data on temporal behavior; this is an area of work in the near future.

6. CONCLUSIONS AND FUTURE WORK

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 more than a million users and a hundred million calls, over different time periods. We found evidence suggesting that key distributions (of the per-user number of distinct call partners, number of calls and number of minutes) have fundamentally different characteristics from power-law and lognormal distributions. Using our proposed method *PowerTrack*, we found significantly better fits using the DPLN distribution. DPLN generalizes the power-law and lognormal distributions using four parameters that can be easily monitored. We found that these parameters remained stable over time in our datasets.

We also found that our graph evolved over time in a way consistent with a generative process based on geometric Brownian motion. Furthermore, this generative process lends itself to a natural and appealing *social wealth* interpretation, giving a plausible reason for the success of *PowerTrack* , and also allowing for extrapolations and interpolations. We hope that our success with *PowerTrack* spurs further studies involving other datasets and their underlying generative processes. In particular, we hope that our analysis will serve as an incentive to study the large-scale evolutionary aspects of social characteristics.

7. REFERENCES

[1] L. A. Adamic and B. A. Huberman. The Web's

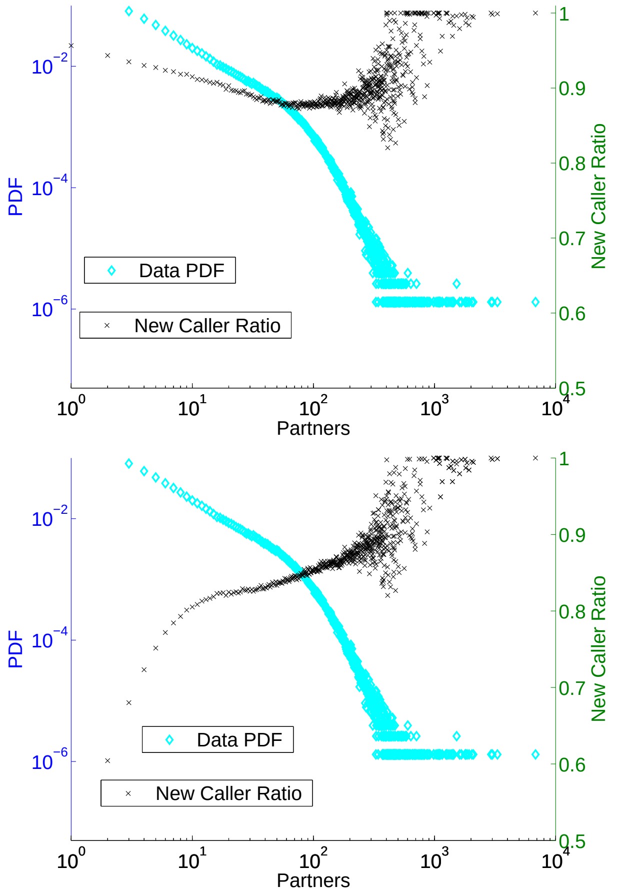

Figure 8: Distribution of η across $T1, T_2$ for all users at Switch S_1 (Top), and for only users active during both T_1, T_2 (Bottom).

hidden order. *Communications of the ACM*, 44(9):55–60, 2001.

[2] L. A. N. Amaral, A. Scala, M. Barthélémy, and H. E. Stanley. Classes of small-world networks. *Proceedings of the National Academy of Sciences*, 97(21):11149–11152, 2000.

[3] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[4] Z. Bi, C. Faloutsos, and F. Korn. The DGX distribution for mining massive, skewed data. In *Proceedings of ACM KDD*, pages 17–26, New York, NY, 2001. ACM Press.

[5] P. Boldi, B. Codenotti, M. Santini, and S. Vigna. Structural properties of the African Web. In *International World Wide Web Conference*, New York, NY, 2002. ACM Press.

[6] A. Z. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web: experiments and models. In *International World Wide Web Conference*, New York, NY, 2000. ACM Press.

[7] A. Clauset. Power law distributions in empirical data. `http://www.santafe.edu/~aaronc/powerlaws/`

[8] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *ArXiv e-print 0706.1062v1*, 2007.

[9] C. Cortes, D. Pregibon, and C. Volinsky. Communities of interest. In *IDA '01: Proceedings of the 4th International Conference on Advances in Intelligent Data Analysis*, pages 105–114, London, UK, 2001. Springer-Verlag.

[10] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the Internet topology. In *Proceedings of ACM SIGCOMM*, pages 251–262, New York, NY, 1999.

[11] R. Govindan and H. Tangmunarunkit. Heuristics for Internet map discovery. In *IEEE INFOCOM*, pages 1371–1380, Los Alamitos, CA, March 2000. IEEE Computer Society Press.

[12] J.-P. Onnela, J. Saramaäki, J. Hyvöven, G. Szabó, M. Argollo de Menezes, K. Kaski, and A.-L. Barabási. Structure and Tie Strengths in Mobile Communication Networks. *New Journal of Physics*, 9, 2007.

[13] S. Keshav. Why cell phones will dominate the future Internet. *Computer Communications Review*, 35(2), April 2005.

[14] S. R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Extracting large-scale knowledge bases from the web. *VLDB*, pages 639–650, 1999.

[15] M. Mitzenmacher. A Brief History of Generative Models for Power Law and Lognormal Distributions. *Internet Mathematics*, 1(2):226–251.

[16] M. Mitzenmacher. Dynamic Models for File Sizes and Double Pareto Distributions. *Internet Mathematics*, 1(3):305–334, 2004.

[17] A. A. Nanavati, S. Gurumurthy, G. Das, D. Chakraborty, K. Dasgupta, S. Mukherjea, and A. Joshi. On the structural properties of massive telecom call graphs : findings and implications. *Proc. of 15th ACM Conference on Information and Knowledge Management*, pages 435–444, 2006.

[18] M. E. J. Newman. Power laws, pareto distributions and Zipf's law. *Contemporary Physics*, 46:323–351, 2005.

[19] G. Palla, A.-L. Barabasi, and T. Vicsek. Quantifying social group evolution. *Nature*, 446(664), 2007.

[20] V. Pareto. *Oeuvres Completes*. Droz, Geneva, 1896.

[21] D. M. Pennock, G. W. Flake, S. Lawrence, E. J. Glover, and C. L. Giles. Winners don't take all: Characterizing the competition for links on the Web. *Proceedings of the National Academy of Sciences*, 99(8):5207–5211, 2002.

[22] S. Redner. How popular is your paper? an empirical study of the citation distribution. *The European Physics Journal B*, 4:131–134, 1998.

[23] W. Reed and M. Jorgensen. The double pareto-lognormal distribution - a new parametric model for size distribution. *Communications in Statistics -Theory and Methods*, 33(8):1733–1753, 2004.

[24] R.Gibrat. *inégalités économiques. Librarie du Recuil Sirey*, 1931.

[25] G. Szabo and A.-L. Barabasi. Network effects in service usage. *ArXiv e-prints physics/0611177*, November 2006.