On the random nature of MySpace friendships

UMass Technical Report UM-CS-2009-021

Bruno Ribeiro[1], William Gauvin[2], Benyuan Liu[2], and Don Towsley[1]

[1]Computer Science Department
University of Massachusetts Amherst
Amherst, MA, 01003
{ribeiro, towsley}@cs.umass.edu

[2]Computer Science Department
University of Massachusetts Lowell
Lowell, MA 01854
{wgauvin,bliu}@cs.uml.edu

ABSTRACT

In this work we study the MySpace friendship graph and provide strong evidence that short account activity lifespans are directly responsible its (truncated) power-law degree distribution. The activity lifespan is the time elapsed since the creation of the account until the user's last login time. Our observations from accounts that have the same lifespans show their degree distribution to be well approximated by a lognormal (with a fairly light tail). Using the Central Limit Theorem we argue that even if MySpace users independently added friends at random we would still observe the same degree distributions. The key insight in our work is to look at the social graph as a live system where users can add or remove friends only when they are active.

1. INTRODUCTION

On-line social network (OSN) friendship graphs have been the subject of a large body of work in the literature complex networks. Arguably, the most studied metric of such graphs is their degree distribution. In this work we randomly sample nearly 400,000 MySpace accounts and show that many users have very short activity lifespans. The activity lifespan is the time elapsed since the creation of the account until the user's last login time. This short lifespans have a profound impact on the degree distribution. Our data provides evidence that the double-Pareto (refer to [12] for the definition of double-Pareto) shaped degree distribution observed in MySpace [3] is a consequence of a mixture of lognormal distributions with exponentially distributed activity lifespans.

MySpace is one of the largest on-line social networks to date. MySpace has approximately 200 million accounts (users) geographically distributed around the globe. The choice of MySpace for our study comes from two valuable records available in most of MySpace accounts: The date in which the account was created and the user's last login date. By randomly sampling MySpace accounts we observe that:

- The distribution of account activity lifespans decays at least as fast as an exponential.
- The friendship degree is lognormally distributed within accounts with the same lifespan. We comment on the implications that this could be a direct consequence of the Central Limit Theorem.
- The friendship graph degrees follow a (truncated) double-Pareto distribution. Moreover, borrowing from Reed [12], we argue that this heavy tailed distribution is a direct consequence of the previous two observations above.

This work differs from previous OSN works in that we look at the social graph as a live system where users can add or remove friends only when they are active. This work is also exploratory in nature. We answer a number of questions but we also leave many others open.

This work is organized as follows. In Section 2 we review the data collected from MySpace. In Section 3 we analyze the data collected from MySpace. Section 4 uses the Central Limit Theorem in order to show that it is possible to obtain the degree distributions seen in Section 3 assuming that MySpace users independently add friends at random. We show that a simple stochastic process can approximate the degree distributions without making assumptions on how friends connect to each other.

2. MEASUREMENT METHODOLOGY

Unfortunately, studying such a large and active social network has its drawbacks. The massive number of users combined with MySpace's stringent rules on crawling its network forces researchers to rely on statistics from incomplete datasets. We collect data from MySpace by sampling, uniformly at random, user profiles and their *blog* entries. An entry in our dataset is comprised of user ID, IDs of all his or her friends, the date in which the account was created, and the user's last login date. Our data was collected using two probing phases: In the first phase, denoted "fast probing", obtains a (time) snapshot of the MySpace graph. In 4

days we randomly sampled 1 million IDs where 70,000+ correspond to **valid** public accounts. This measurement had to be shut down due to complaints from MySpace. In the second phase, denoted "slow probing", we obtain 312,713 **valid** public MySpace accounts, chosen uniformly at randomly in the account space, during 7 months of measurements.

The data collected in the "fast probing" phase is used for our snapshot-sensitive analysis, e.g. the lifespan distribution. As we are not too interested in the tail of these distributions, we hypothesize that 70,000+ samples are enough to obtain good estimates. The data collected in the "slow probing" phase is used for all snapshot-insensitive analysis such as the distribution of friends from accounts with a given lifespan. The results obtained from the fast probing phase is also used to double check the results obtained in the slow probing phase. In this preliminary work we hypothesize that accounts which are closed before the network is sampled do not interfere with our conclusions. We also hypothesize that private profiles, also reported in [3], (from which we cannot obtain friends information) do not affect our results. We leave as future work the task to collect data that can verify these hypotheses.

One of the challenges of this work is to perform statistical analysis using relatively few samples. The quality of our conclusions depends directly on the quality of our estimates. In our experiments we sample nearly 0.25% of all valid users. In Appendix A we analyze the impact of the incomplete data over our estimates. In what follows we describe the statistics obtained using this data.

3. MYSPACE GRAPH DEGREES

In this section we look at the impact of lifespans over the MySpace friendship graph. The friendship graph $G = (V, E)$ is an undirected graph where vertices are MySpace accounts. Two accounts u and v have an edge in G if u and v are friends in MySpace. In what follows we look at account lifespan and vertex degree distributions. While the vertex degree distribution of social graphs has been extensively studied in the literature, including a MySpace study [3], we show crucial statistical properties that have escaped the attention of previous works.

3.1 Measures of graph growth an activity

We introduce three statistics that measure the growth (in terms of new accounts) and how active MySpace users are:

- **Account lifespan:** Time between the creation of an account and the last time the user logged in.
- **Account age:** Time between the creation of an account and when it is probed (recorded in our trace).
- **Account inter-login time:** Time between two consecutive logins into the same account.

These statistics have a deep impact on how the friendship graphs grows and its vertex degree distribution. Figure 1 shows the complementary cumulative distribution function (CCDF) of MySpace lifespans where the y-axis is shown in log scale. We see that the majority of users in MySpace are active for a very short period of time. The CCDF of lifespans can be divided into two parts. The first part with lifespans < 26.5 months follow an exponential distribution (straight line in log-scale). This first part accounts for more than 80% of the accounts. The second part with lifespans ≥ 26.5 months follow a parabola ($\exp(-\text{lifespan}^2)$) in log-scale. The fast tail decay is, in part, a consequence of the truncation of the distribution, as MySpace was launched in August 2003 and the data was collected in March 2009 (65 months later). The shaded area in Figure 1 shows the fitted distributions and their divisions.

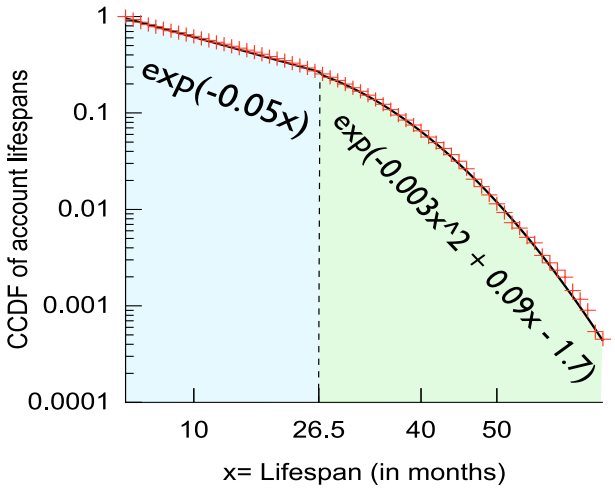

Figure 1: Empirical complementary cumulative histogram of user activity lifespans. The red points represent the lifespan distribution observed in our data and the lines correspond to the curves shown in the equations below the curve. More than 80% of the probability mass follows an exponential law (the remaining 20% decays faster than an exponential).

Figure 1 may leave the false impression that exponential lifespans are a direct consequence of the exponential growth of the graph, i.e., lifespans are exponentially distributed because account ages are exponentially distributed. This is not the case for MySpace. Figure 2 shows the distribution of account ages (in months). Note that, unlike lifespans, at least 80% of the MySpace accounts (accounts newer than 37 months) are the result of linear growth rates. Accounts older than 37 months

are the part of the graph that grew exponentially during MySpace's early years. Less than 20% of the vertices were created during this exponential growth period.

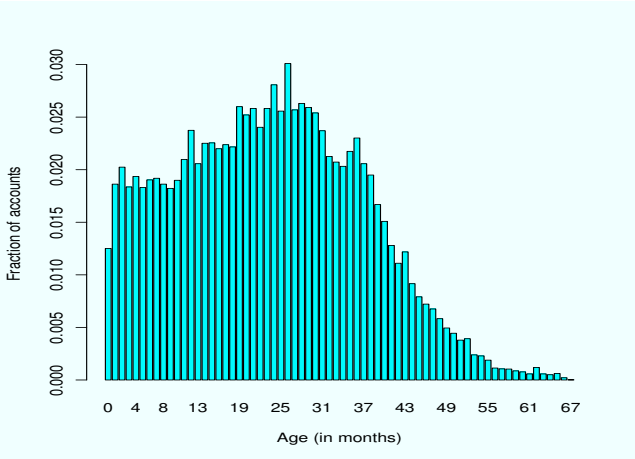

Figure 2: Fraction of MySpace accounts with age = (<Time of scan> - <Member Since>). After an exponential growth from 2003 (MySpace's launch) to 2005, the number of new accounts transitions to linear growth.

We believe that account lifespans are one of the most important statistics that one can obtain from an OSN such as MySpace. We also argue that account ages are not as relevant. This is because friends are not automatically added in MySpace. Users must log into their accounts in order to add friends. Therefore, an account created and later abandoned cannot play a significant role in the graph evolution after it is abandoned. We can sample such accounts because MySpace never deletes accounts due to inactivity.

The lifespan distribution brings us to another question: How frequently do users log into their accounts? Note that one could generate the same lifespan statistics if users logged in just once. In order to answer this question we need to estimate the time between two consecutive logins into the same account (*inter-login times*). Assuming that our probes arrive at points in time that are distributed uniformly at random, we can estimate inter-login times using the account's last login time and the time of the probing. It is clear that we are more likely to probe long inter-login times than short ones. This is known as the *inspection paradox*. Appendix B presents a maximum likelihood estimator that is used to obtain the graph in Figure 3. In our estimates we artificially include a constraint that there are no inter-login times greater than 3 years in order to speed the computation. Unfortunately, we can only rely on our estimates as we do not have access to the ground truth. However, the results shown in Figure 3 seem to agree with our intuition. First, we observe that most accounts are logged in quite frequently. This is

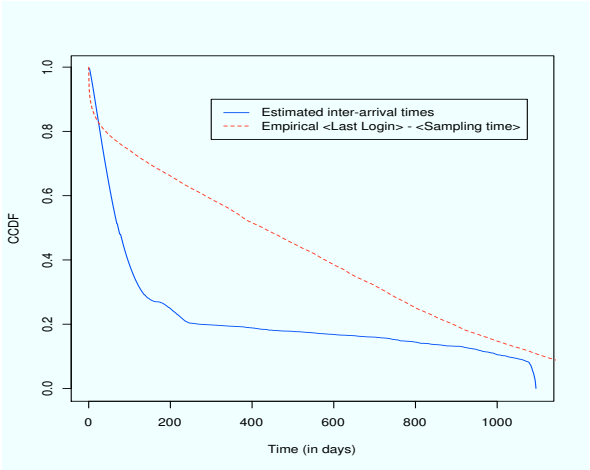

Figure 3: CCDF of estimated inter-login times. Note that a heavy tail is expected as many users abandon their MySpace accounts. The sharp drop at the end of the tail is due to an artificial constraint that there are no inter-login times greater than 3 years.

a sign that users login quite often during the account lifespan. Also, most accounts that are not active in the span of one year are not likely to be active in less than three years. This is expected as accounts inactive for more than one year are likely to have being abandoned. Figure 3 also shows the distribution obtained from the difference between the time of the probing and the account's last login time, which is the input data used in our estimator.

3.2 Conditional friend degree distribution

Another important statistic missing from the literature is the degree distribution of accounts with the same lifespan, i.e., the degree distribution conditioned on the lifespan. Figure 4 shows QQ-plots that test if the degree distributions of accounts with 3 to 65 months of lifespan are lognormally distributed. Figure 4(a) shows all 63 QQ-plot curves. The straight line is the perfect match to a lognormal distribution. Because many curves in Figure 4(a) intersect, we opt to also show, in Figure 4(b), the heatmap of Figure 4(a) where colors (from blue to yellow) indicate the density of overlapping points (from low to high, respectively). From these graphs we see that all these distributions can be well described by a lognormal law. Note that both axes in Figure 4 are not the number of friends as seen in a regular QQ-plot. This is an artifact we use to get around the fact that accounts with different lifespans have different lognormal parameters. If plotted in their regular scale, these curves are not comparable. We instead apply a simple transformation observing that the log of a lognormal random variable is Normally distributed: Apply the log to the data, subtract the result

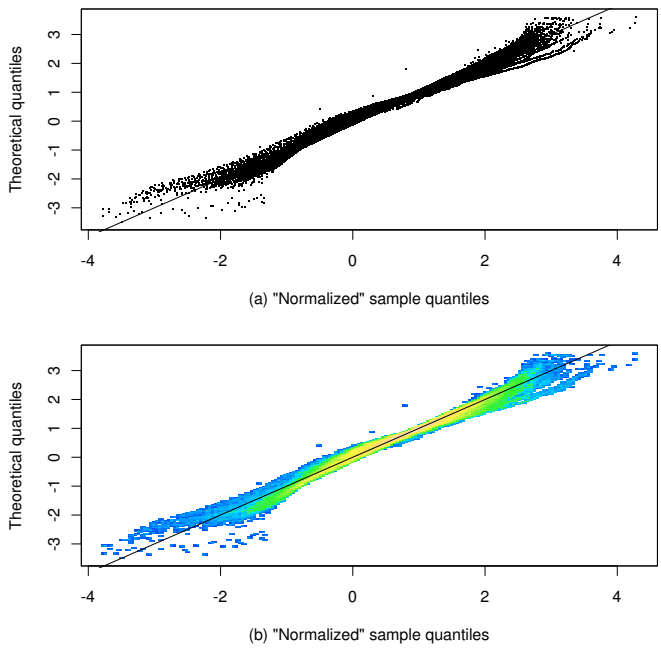

Figure 4: QQ-plots of the degree distributions from accounts with the same number of months of activity lifespan. These graphs plot 63 curves that correspond to the degree distributions of 3 months of activity until up to 65. The theoretical quantiles are given by the t-Student distribution (tests if the samples come from a standard Normal).

from their sample average, and divide it by the sample standard deviation. Thus, if the original data is lognormal, the new transformed ("normalized") data must be distributed according to a t-Student distribution whose degrees of freedom is the number of data points. In the graphs of Figure 4 we see that these distributions can be well described by a lognormal law. The only three

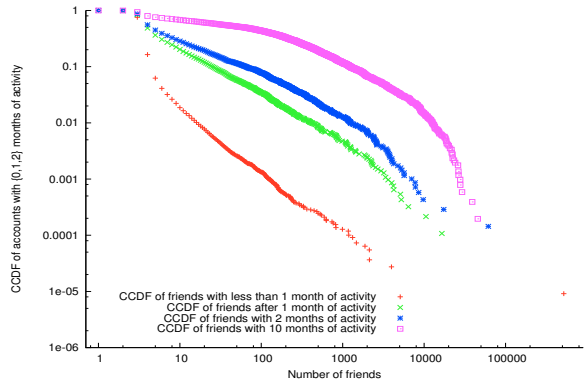

Figure 5: Log-log plot of the CCDF of friends for accounts with lifespans 0 (less than one month), 1, 2, and 10 months.

distributions that are not well described by a lognormal are (unsurprisingly, as seen in Section 4) the three degree distributions from accounts that have less than two months of activity. These three distributions are shown in Figure 5 along with a "typical" distribution among the ones analyzed in the QQ-plot of Figure 4 (10 months of activity). In Figure 5 we see that the distributions from accounts with short lifespans look closer to a power-law than a lognormal-law. Section 4 presents one possible mathematical reason behind this phenomenon. Another likely cause are bots (programs that automatically send friend requests to other MySpace users from bogus accounts). MySpace closely monitors its users. If an user behaves suspiciously, MySpace blocks the account until the user proves to be legitimate. Thus, we expect to find some user accounts with short lifespans and a large number of friends.

Estimates of the lognormal parameters (μ, σ) for each lifespan value (in months) can be found in the graph of Figure 6. Note that parameter μ seems to grow linearly with T for lifespans greater than 4. On the other hand, parameter σ seems to be constant for lifespans greater than 4. In what follows we combine the above results to understand the graph degree distribution of MySpace.

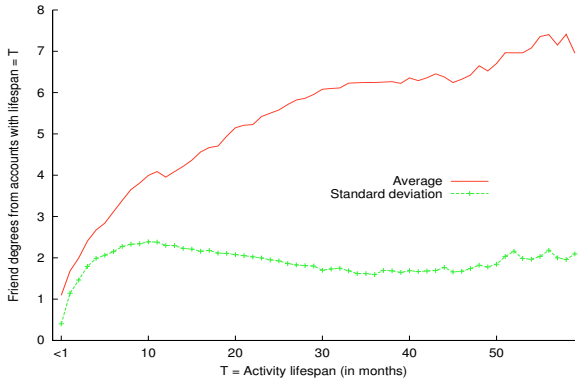

Figure 6: Empirical average and standard deviation of friend degree given account lifespan T shows a linear increase for $T > 10$ while the standard deviation remains constant.

3.3 Assembling the puzzle: The friend degree distribution

Reed [12] shows that the convolution of lognormally distributed random variables with parameters $(T\mu, T\sigma^2)$ given a fixed lifespan T with T itself being exponentially distributed results in a distribution that can be approximated by a double-Pareto distribution. The double-Pareto distribution is characterized by its graph in log-log scale: Two straight lines connected through a "knee". The friend degree distribution graph of MySpace shown in Figure 7 is an excellent example of this shape. The same shape is also present in a previous measure-

ment study of MySpace [3].

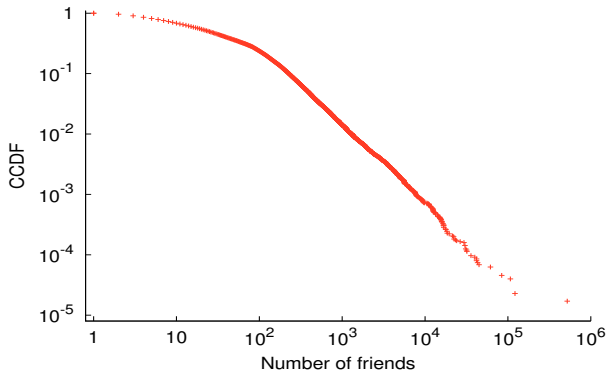

Figure 7: Empirical Complementary cumulative distribution of the number of friends in MySpace accounts.

The double-Pareto shape of MySpace's degree distribution is not surprising according to the statistics seen in this section as the degree distribution is a convolution of:

1. Exponential lifetimes (Figure 1) (where active users log into their accounts quite frequently, as seen in Figure 3) with
2. the lognormal distribution of friends given a fixed lifespan (Figure 4) with parameters $(T\mu, \sigma^2)$ (Figure 6).

The standard deviation of the lognormals is the main difference between the statistics presented in this section and Reed's model [12]. In Reed's model the standard deviations are $\sigma\sqrt{T}$ whereas in MySpace they are just σ. While this is not a trivial difference, the resulting distribution in this case is also double-Pareto shaped as seen in Figure 7.

3.4 Related work

Closely related to the above observation is the observation of Huberman and Adamic [5], in 1999, that the exponential growth of the World Wide Web (WWW) graph could explain its power law degree distribution. A webpage, like a MySpace user, adds and removes links ("friends"). But note that the model in Huberman and Adamic [5] implicitly assumes that most webpages undergo sustained changes (addition and deletion of links) from the moment they were created until when the page is sampled. This is equivalent to assume that webpages are never abandoned. While this is a fair assumption about the WWW in 1999, this assumption does not apply to MySpace, as many MySpace users create accounts and quickly abandon them. In MySpace, Huberman and Adamic's assumption of exponential graph growth is replaced by the exponential lifespans (during which MySpace users are able to include and remove friends). Mitzenmacher [11] has proposed a mechanism similar to Reed's to describe Web file sizes. Seshadri et al. [13] has proposed a similar mechanism to describe the duration of cell phone calls which makes assumptions about the wealth of the callers. Different from the above works, we are able to empirically verify all the required conditions to generate a double-Pareto distribution. In what follows we provide a model that generates the statistics seen in this section but does not make assumptions on how users connect to each other in the graph.

4. MYSPACE DEGREE DISTRIBUTIONS AND THE CENTRAL LIMIT THEOREM

It is well known in the literature of complex networks that there are many possible generative graph models that entail graphs with the same heavy-tailed degree distributions [9]. The literature has many examples of these generative graph models. The reader finds in Liu et al. [8] a good reference for the connection between graph degree distributions, generative models, and heavy-tails. Also, we can refer the reader to Mitzenmacher [10] for a survey of generative models specific to power-law distributions.

In this section we do **not** propose a new model to generate the entire MySpace graph. Rather, we provide a model (stochastic process) that approximates the degree distributions seen on MySpace and does not need assumptions on how users connect to each other in the graph. The following model is similar to the one proposed by Huberman and Adamic [5] for the degree distribution of the WWW graph. Let X_d be a random variable that denotes the number of friends of a randomly chosen user with lifespan of d days. Our model assumes that

$$X_d = F_d X_{d-1} \tag{1}$$

where F_i, $i = 1, 2, \ldots$ are independent random variables with finite mean and variance and $X_0 = 1$ (MySpace users start with "Tom" (MySpace's creator) as their friend). Applying the log to both sides of equation (1) we have

$$\log(X_d) = \sum_{i=1}^{d-1} \log(F_i). \tag{2}$$

The Central Limit Theorem (CLT) states that an infinite sum of independent random variables, where no random variable dominates the sum[1] in equation (2), converges to the Normal distribution [14]. A direct consequence of the CLT is that the Normal distribution

[1]It is easy to see that as the number of elements in sum goes to infinity the assumption that "no random variable dominates the sum" can be replaced by the assumption that each F_i has finite mean and variance.

is stable, i.e., the sum of two Normal distributions is also Normal. The assumption that no random variable dominates the sum is actually more important than the assumption of an infinite sum [14]. So, it is reasonable to expect that if no sample of the $\log(F_i)$'s dominates the sum, even finite sums can be well approximated by a Normal distribution. This simple model provides a plausible reason (given a set of reasonable assumptions) that is able to explain the lognormals seen in the conditional degree distributions in Section 3.2. It is important to note that this model makes no assumptions on how friends connect to each other.

It is easy to see that for small values of d it is likely that one of the samples ($\log(\hat{F}_i)$, $i = 1, \ldots, d-1$) dominates the sum. This is another possible explanation (besides the presence of bots) for the graphs in Figure 5. In fact, it is not surprising that Figure 5 indicates that the distribution from accounts with a 2-month lifespan is much closer to a lognormal than the distribution of accounts with 1 or 0 month lifespans. The model in equation (1) assumes that MySpace users choose friends independently. As the friendship graph in MySpace is undirected, this assumption is not entirely true. But it is reasonable to assume that two randomly chosen users add and remove friends independently. This is enough independence to explain the good lognormal approximations seen in Figure 4. As we have seen in Section 3, the double-Pareto degree distribution of the MySpace friendship graph follows from the exponential lifespans and the linear increase in the average of the lognormals over time. It is fair to say that the multiplicative nature of equation (1) is a type of "preferential attachment" [1]. However, it is worth noting that equation (1) makes no assumption about how the graph is going to be constructed by its agents (in this case, agents are MySpace users).

5. SUMMARY & CONCLUSIONS

In this work we study the MySpace friendship graph and provide strong evidence that short account activity lifespans are the reason behind its (truncated) power-law degree distribution. In Section 3 we see that accounts with the same lifespans have their degree distribution following a lognormal-law (with a fairly light tail). In Section 4 we use the Central Limit Theorem in order to show that it is possible to obtain these degree distribution even if MySpace users independently added friends at random. We show that a simple stochastic process can approximate the degree distributions without making assumptions on how friends connect to each other.

APPENDIX

A. THE IMPACT OF SAMPLING ON OUR ESTIMATES

This section is dedicated to explain the methodology used to substantiated our claims and describe the implications of working with incomplete (sampled) data. The following exposition is quite straightforward but needed to ensure us that our conclusions are sound. Fitting distributions to sampled data is somewhat of a controversial topic [4]. In the complex networks literature heavy-tailed distributions are often found in observed (incomplete) data: links in Web pages [1, 5], file sizes [11], among many others. This comes as no surprise as, according to the theory of stable laws, heavy-tailed distributions are easily generated from a number of stochastic processes. In what follows we analyze the estimation error and the maximum likelihood estimate for our data.

A.1 Truncated tail

The study of heavy-tailed distributions requires a brief warning about the tail of the distribution. In most, if not all, scenarios these tails are truncated. A good example is the distribution of the energy of earthquakes. While, from the sampled data already collected, such distribution appears to be heavy-tailed, it is clear that the tail of the distribution is not truly "heavy" as there is a limit to the amount of energy that can be released from the Earth's interior [7]. Our application is no exception and has an obvious truncation point (the number of users in MySpace). In what follows we refer to the "tail" of our distributions as all points that are "far from zero" but smaller than the truncation point. While there is great inaccuracy in measuring the tail [4], and our measurements are no exception, there is still much that can be said about the tail. In what follows we show how this is possible.

A.2 Estimation error

Let θ_i be the fraction of MySpace accounts with i friends and $\boldsymbol{\theta} = \{\theta_i | i = 1, \ldots\}$. Let

$$\Theta_d = \sum_{i=d+1}^{\infty} \theta_i,$$

be the fraction of accounts with more than d friends. Let $\mathbf{Y} = \{Y_i\}_{i=1}^{N}$ be the (incomplete) raw data obtained from N sampled MySpace accounts. We define the *sampled distribution* to be the distribution of friends obtained from the incomplete dataset $\mathbf{Y}$. Let $T_d(\mathbf{Y})$ be an unbiased estimate of Θ_d, i.e., $E[T_d(\mathbf{Y})] = \Theta_d$. Also let

$$T_d^{\star} = \operatorname*{argmin}_{T_d} E[(\Theta_d - T_d(\mathbf{Y}))^2],$$

i.e., estimator $T^{\star}$ has the smallest mean squared error

among all unbiased estimators (we assume Hajék regularity [6]). In what follows we answer the following questions:

(1) How accurate can $T^\star$ be?

(2) What is the most likely shape of the original distribution?

For now we assume that the only information about $\boldsymbol{\theta}$ contained in the accounts is the number of friends. If accounts display only the number of friends (not their IDs) and as we sample accounts independently at random, sampling accounts is equivalent to sampling degrees directly from $\boldsymbol{\theta}$. Let $\#(\mathbf{Y} == d)$ denote the number of sampled accounts with d friends. We have

$$P[\#(\mathbf{Y} == d) = k] = \theta_d^k (1-\theta_d)^{N-k}.$$

From the above it is easy to see that the following inequality holds

$$E\left[(\Theta_d - T(\mathbf{Y}))^2\right] \geq \frac{\Theta_d(1-\Theta_d)}{N}. \quad (3)$$

The above inequality is a straightforward application of the Cramér-Rao inequality [2]. And $T^\star$ obtains the sampled distribution making the bound in eq. (3) tight. The above analysis is quite trivial but it has a remarkable impact on our ability to draw conclusions from our sampled data. In order to exemplify the implications of equation (3) over the accuracy of our estimates, we assume, for the sake of argument, that $\Theta_d = d^{-\mu}$ with $\mu \geq 1$ and $d = 1, \ldots,$ i.e., distribution $\boldsymbol{\theta}$ is Pareto with scale=1 and shape $\mu \geq 1$. The empirical distribution of the number of friends in our MySpace traces has shape parameter $\hat{\mu} = 1.47$ at the tail. In order to simplify our analysis we use another metric of accuracy, the normalized root mean squared error (or NRMSE):

$$\text{NRMSE}(T) = \frac{\sqrt{E\left[(\Theta_d - T(\mathbf{Y}))^2\right]}}{\Theta_d} \geq \sqrt{\frac{d^\mu - 1}{N}},$$

recall that N is the number of sampled MySpace accounts. The inequality in the equation above comes from equation (3). The NRMSE is a metric that gives the average error of estimator T as a fraction of the quantity being estimated. With $N = 4 \times 10^5$ (400,000 sampled accounts) we have

$$\text{NRMSE}(T) \geq \sqrt{(d^\mu - 1)/(4 \times 10^5)}.$$

Thus, any unbiased estimate of $\Theta_{10,000}$ has an average NRMSE of at least $1.38 \cdot \Theta_{10,000}$. This reasonably large error is one of the reasons why fitting a distribution to the tail of a sampled distribution is a controversial topic. Please refer to Gong et al. [4] for an interesting look at the difficulty in estimating the tail of Web file size distributions. An interesting question for future work is whether the poor accuracy of $T^\star$ implies that we cannot be confident about the shape of the original distribution. In what follows we estimate the likely shape of the distribution.

A.3 Maximum likelihood estimation

Here we show the trial task of estimating the likely shape of our sampled distributions. We look at the distribution that most likely generated the data finding the most likely non-parametric distribution that generated the sampled data. We wish to find $\hat{\boldsymbol{\theta}}$ that maximizes the probability that $\mathbf{Y} = \mathbf{y}$, i.e., we wish to find the maximum likelihood estimate

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmax}}\, P[\mathbf{Y} = \mathbf{y} \,|\, \boldsymbol{\theta}].$$

The maximum likelihood estimate of the above Binomial random variable is $\hat{\theta}_d = 1/\#(\mathbf{y} == d)$, i.e., $\hat{\boldsymbol{\theta}}$ is the sampled distribution of $\mathbf{y}$. Note that $\#(\mathbf{y} == d)$ denotes the number of sampled accounts with d friends. Figure 7 shows the empirical distribution of friends in our MySpace traces. The rather trivial conclusion is that the empirical curve seen in Figure 7 is the most likely distribution of friends in MySpace.

A.4 Changes to the original (incomplete) data

Zero friends: When someone joins MySpace they have the creator of MySpace "Tom" as their friend. Thus, there are very few accounts with zero friends. For the sake of simplicity we ignore accounts with zero friends.

B. USER INTER-LOGIN TIME DISTRIBUTION

Let Y be the time (in days) between when an account is probed and the last time it was logged in. If the difference in time is less than 24 hours then $Y = 1$, if the difference in time is between 24 and 48 hours then $Y = 2$, and so forth. Assume that, collectively, users login an infinite number of times. Let X be the time (in days) between two consecutive logins of an user. In what follows we assume that accounts do not go stale (in reality many users abandon their accounts).

If we assume that the time we sample the account is distributed uniform at random, the probability of landing on an interarrival time of size x is

$$P[Y = i | X = j] = \begin{cases} 0 & \text{if } j < i \\ 1/i & \text{otherwise} \end{cases}. \quad (4)$$

The probability that we will sample an interval X of size j is

$$\frac{jP[X = j]}{\sum_{k=1}^{\infty} kP[X = k]}. \quad (5)$$

Putting equations (4) and (5) together we have

$$P[Y=i] = \sum_{j=i}^{\infty} \frac{1}{j} \frac{jP[X=j]}{\sum_{k=1}^{\infty} kP[X=k]} = \frac{P[X \geq i]}{E[X]}$$

Thus we can recursively calculate $P[X \geq i]$ from:

$$\begin{aligned} E[X] &= \frac{1}{P[Y=1]}, \quad \text{and ,} \\ P[X \geq i] &= E[X]P[Y=i]. \end{aligned}$$

As we only have an estimate of $P[Y=i]$ and not its true value, the above estimate is subject to sampling noise. Indeed, using the above estimator in our dataset we obtain a number of negative $P[X=j]$ values. In order to obtain better estimates, we use the maximum log-likelihood estimator

$$\underset{\{P[X=j]\}}{\operatorname{argmax}} \sum_{\forall i} y_i \frac{1 - \sum_{j=1}^{i-1} P[X=j]}{\sum_{k=1}^{\infty} kP[X=k]},$$

where y_i is the number of samples of Y with value i. We also enforce the constraints $0 \leq P[X=j] \leq 1$, $j = 1, 2, \ldots$ and $\sum_{\forall j} P[X=j] = 1$.

1. REFERENCES

[1] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, October 1999.

[2] George Casella and Roger Berger. *Statistical Inference*. Duxbury Resource Center, June 2001.

[3] James Caverlee and Steve Webb. A large-scale study of MySpace: Observations and implications for online social networks. In *Proceedings from the 2nd International Conference on Weblogs and Social Media (AAAI)*, 2008.

[4] Weibo Gong, Yong Liu, Vishal Misra, and Don Towsley. On the tails of Web file size distributions. In *Proceedings of 39-th Allerton Conference on Communication, Control, and Computing*, Oct. 2001.

[5] Bernardo Huberman and Lada Adamic. Growth dynamics of the World-Wide Web. *Nature*, pages 130–130, 1999.

[6] I.A. Ibragimov and R.Z. Khasminskii. *Statistical Estimation. Asymptotic Theory*. Springer, New York, 1981.

[7] L. Knopoff and Y. Kagan. Analysis of the theory of extremes as applied to earthquake problems. *Journal of Geophysical Research*, 82:5647–5657.

[8] Lun Li, David Alderson, John Doyle, and Walter Willinger. Towards a theory of scale-free graphs: Definition, properties, and implications. *Internet Mathematics*, 2(4), 2005.

[9] Lun Li, David Alderson, John C. Doyle, and Walter Willinger. Towards a theory of scale-free graphs: Definition, properties, and implications. *Internet Mathematics*, 2005.

[10] Michael Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1(2), 2003.

[11] Michael Mitzenmacher. Dynamic models for file sizes and double pareto distributions. *Internet Mathematics*, 1(3), 2003.

[12] William J. Reed. The pareto, zipf and other power laws. *Economics Letters*, 74(1):15–19, December 2001.

[13] Mukund Seshadri, Sridhar Machiraju, Ashwin Sridharan, Jean Bolot, Christos Faloutsos, and Jure Leskove. Mobile call graphs: beyond power-law and lognormal distributions. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 596–604, New York, NY, USA, 2008. ACM.

[14] Didier Sornette. *Critical Phenomena in Natural Sciences: Chaos, Fractals, Selforganization and Disorder: Concepts and Tools*. Springer, April 2006.