On MySpace Account Spans and Double Pareto-Like Distribution of Friends

UMass Technical Report UM-CS-2010-001

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 activity span of MySpace accounts and its connection to the distribution of the number of friends. The activity span is the time elapsed since the creation of the account until the user's last login time. We observe exponentially distributed activity spans. We also observer that the distribution of the number of friends over accounts with the same activity span is well approximated by a lognormal with a fairly light tail. These two findings shed light into the puzzling (yet unexplained) shape of the distribution of friends in MySpace when plotted in log-log scale: Two straight lines with different parameters joined by an inflection point (knee). We argue that the inflection point is more likely related to the inflection point of Reed's (Double Pareto) Geometric Brownian Motion with Exponential Stopping Times model than to the Dunbar number hypothesis of online social networks, which argues, without proof, that the inflection point is due to the Dunbar number (a theoretical limit on the number of people that a human brain can sustain active social contact with). While we answer many questions, we leave many others open.**

I. Introduction

MySpace is one of the largest on-line social networks to date with approximately 200 million accounts (users) geographically distributed around the globe. In this work we collect 400,000 randomly sampled MySpace accounts. An unbiased estimate of the distribution of the number of friends of MySpace users can be seen in log-log scale in Figure 1. The shape of the distribution seen in Figure 1 agrees with previous unbiased estimates for MySpace [4]. Our findings in this work shed light into the puzzling (yet unexplained) shape of the distribution of friends seen in Figure 1: two Pareto-shaped distributions with different parameters joined by an inflection point (knee). 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: (1) Using this data we find activity spans to be exponentially distributed. This phenomenon may explain much of the shape of the friends distribution shown in Figure 1. The activity span is the time elapsed since the creation of the account until the user's last login time. (2) Inspired by previous works on the Double Pareto distribution [10], [11] we also observe that the distribution of the number of friend among accounts with roughly the same activity span can be well described by a lognormal distribution whose average number of friends grows according to the square root of the account's span. Our findings provide a sound alternative hypothesis for the emergence of a "knee" in the distribution of friends in on-line social networks: The mixture of multiple lognormal distributions with exponentially distributed averages. By noting that no such "knee" exists in the distribution of friend over accounts with roughly the same activity span, we challenge an alternative hypothesis (formulated in [1] for the CyWorld on-line social network) in which the "knee" is seen as a consequence of the Dunbar's number, a theoretical cognitive limit of the number of people (about 147.8) with whom one can maintain stable social relationships [5]. Our work is exploratory in nature. We answer a number of questions, but also leave a number of other interesting questions open.

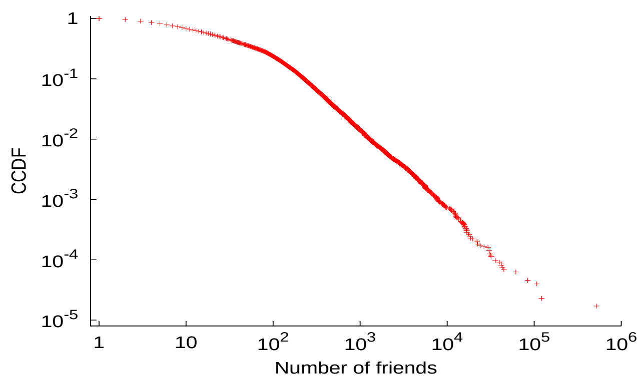


Fig. 1. Empirical Complementary cumulative distribution of the number of friends in MySpace.

This work is organized as follows. In Section II we review the data collected from MySpace. In Section III we analyze the data collected from MySpace. Section IV connects the Double Pareto distribution with the analysis provided in Section III,

This work was supported by ARO MURI W911NF-08-1-0233.

showing possible connections between the number of friends in a MySpace account and Geometric Brownian Motion.

II. Measurement Methodology

Unfortunately, studying such a large and active social network has its drawbacks. The massive number of user accounts 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 account profiles uniformly at random. An entry in our dataset is comprised of user ID, IDs of all user friends, the date in which the account was created, and the user's last login date. Our data was collected in two phases: During the first phase, denoted "fast probing", we obtained 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. During the second phase, denoted "slow probing", we obtained 312,713 **valid** public MySpace accounts over a period of 7 months.

The data collected during the "fast probing" phase is used for our snapshot-sensitive analysis, e.g. the activity span distribution. As we are not too interested in the tail of these distributions, we believe that 70,000+ samples suffice to obtain good estimates. The data collected during the "slow probing" phase is used to obtain the distribution of friends of accounts with the same activity span. 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 private profiles (profiles 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 this 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 accounts. Appendix A analyzes the impact of the incomplete data over our estimates. In what follows we describe the statistics obtained using this data.

III. Data Analysis

In this section we focus on the impact of activity spans on the distribution of friends. In what follows we look at account activity span and friends distribution. While the friends distribution of social networks has been extensively studied in the literature, including a MySpace study [4] (that, like our work, presents an unbiased estimate of the distribution of the number of friends), we show crucial statistical properties that have escaped the attention of previous works.

A. Activity span distribution

In this section we analyze three statistics collected in our experiment:

- **Activity span:** Time between the creation of an account and the last time the user logged in.
- **Age:** Time between the creation of an account and when it is probed (recorded in our trace). Age is also studied in [15].
- **Inter-login time:** Time between two consecutive logins into the same account.

Figure 2 shows the complementary cumulative distribution function (CCDF) of MySpace activity spans where the y-axis is shown in log scale. We see that the majority of accounts in MySpace are active for a very short period of time. The CCDF of activity spans divide into two parts. The first part with activity spans < 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 activity spans $\geq$ 26.5 months follow a parabola ($\exp(-$ activity span$^2))$ 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 2 shows the fitted distributions and their divisions.

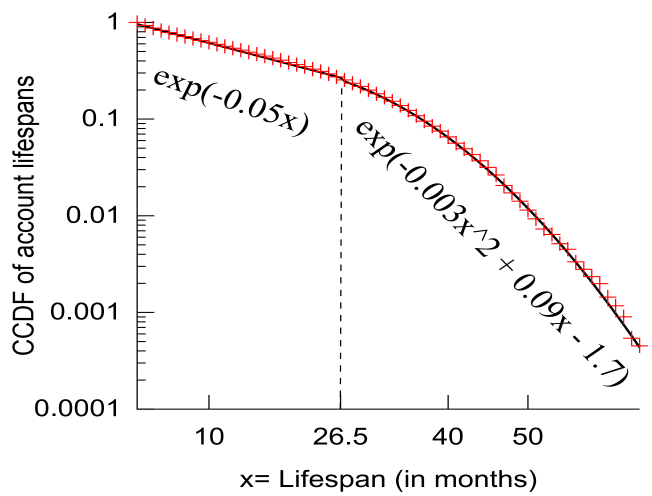


Fig. 2. Empirical complementary cumulative histogram of account activity activity spans. The red points represent the activity span distribution observed in our data and the lines correspond to the curves shown in the equations below the curve. More than 80% of the probability mass follows an exponential law (the remaining 20% decays faster than an exponential).

Figure 2 may leave the false impression that exponential activity spans are a possible consequence of an exponential growth in the number of accounts, i.e., activity spans are exponentially distributed because account ages are exponentially distributed. This is not the case for MySpace. Figure 3 shows the distribution of account ages (in months). Unlike activity spans, the distribution of account ages is not exponentially distributed. We see that less than 20% of the total number of accounts (accounts older than 37 months, created during MySpace's early years) were created when MySpace experienced exponential growth. The remaining 80% of MySpace accounts (accounts newer than 37 months) shows that the recent growth of MySpace has been (at best) linear (the migration of MySpace users has been studied in [15]). These two modes of growth in the age of MySpace accounts may be the reason behind the two distinctive modes in the account activity span distribution.

Fig. 3. 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 activity spans 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 into MySpace accounts. Users must log into their accounts in order to add or accept friends. Therefore, an account created and later abandoned cannot play a significant role in the friends distribution after it is abandoned. Also note that MySpace does not delete accounts due to inactivity.

The activity span distribution brings us to another question: How frequently do users log into their accounts? Note that one could generate the same activity span 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 time*). Assuming that our probes arrive at points in time that are distributed uniformly at random, we can estimate the inter-login time distribution 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 sampling phenomenon is known as the *inspection paradox* [12]. Appendix B presents a maximum likelihood estimator that is used to obtain the CCDF of inter-login times (Figure 4). In order to speedup calculation of our estimates we assume that there are no inter-login times greater than 3 years. We believe this assumption to be reasonable as MySpace had existed for only 6 years at the time of our measurements and we observed that fewer than 20% of MySpace accounts in our trace were created more than 3 years from the time of our measurements. Unfortunately, we can only rely on our estimates as we do not have access to the ground truth. However, the results shown in Figure 4 seem to agree with our intuition. First, we observe that most accounts are logged in quite frequently. This is a sign that users log

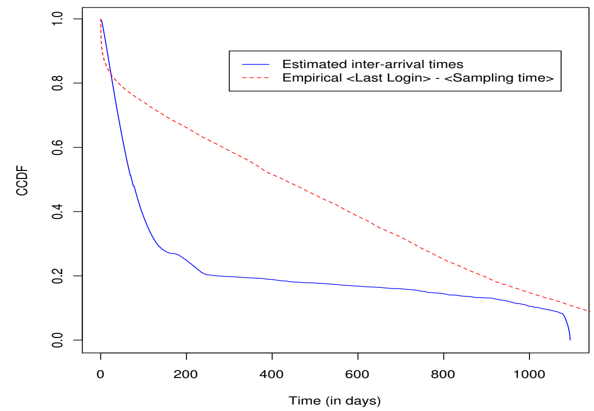


Fig. 4. CCDF of estimated inter-login times. Note that a heavy tail is expected as many users abandon their MySpace accounts. The sharp drop at the end of the tail is due to an artificial constraint that there are no inter-login times greater than 3 years.

in quite often during the account activity span. Also, most accounts that are not active during 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 been abandoned. Figure 4 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.

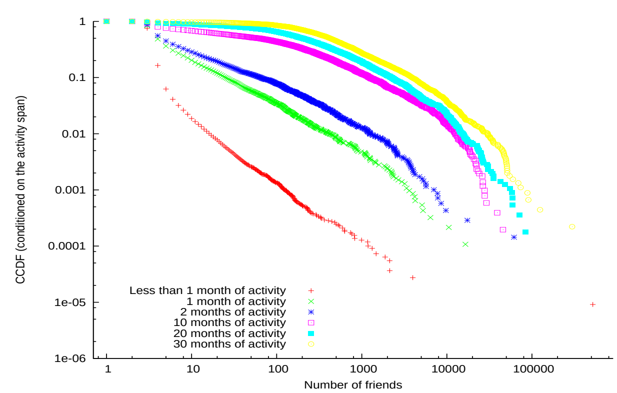
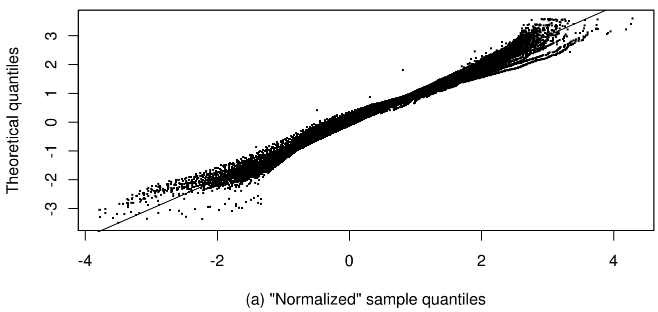


Fig. 5. Log-log plot of the CCDF of friends for accounts with activity spans of $\{<1, 1, 2, 10, 20, 30\}$ months.

B. Conditional friends distribution

Here we present an important statistic missing from the literature: the distribution of MySpace friends from accounts with the same activity span (in months). Figure 5 shows the log-log plot of the CCDF of friends for accounts with activity spans of $\{<1, 1, 2, 10, 20, 30\}$ months. Note that as the activity span increases the CCDF approaches a lognormal shape, the CCDF for activity spans of 10, 20, and 30 months have a shape similar to a lognormal (other months between 3 and 65 have a similar lognormal shape). The CCDF of months



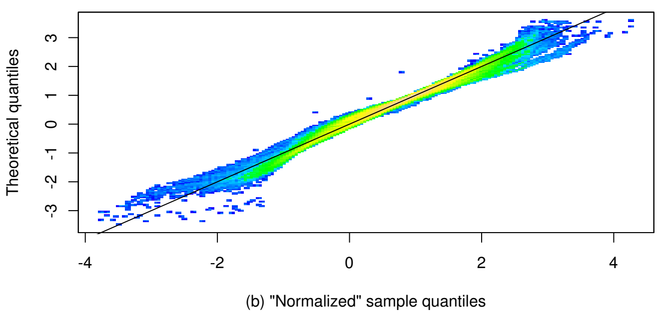


Fig. 6. QQ-plots of the distributions of friends within accounts with the same number of months of activity span. These graphs plot 63 curves that correspond to the distributions of 3 months of activity until up to 65. The theoretical quantiles are given by the t-Student distribution (tests if the samples come from a standard Normal).

$\{1, 2\}$ seem to have an intermediate shape between the shape of < 1 (which is not power-law due to the jump from 3 friends (0.75) to 4 friends (0.16)) and the lognormal shapes of activity spans from 3 to 65 months. The CCDF of activity span < 1 shows an outlier with close to 1 million friends (with less than a month of activity!). A likely cause are bots (programs that automatically send friend requests to other MySpace users from bogus accounts). MySpace closely monitors user accounts. If an user behaves suspiciously, MySpace blocks the account until the user proves to be legitimate. Thus, one is expected to find user accounts with short activity spans and large number of friends. Unfortunately, due to privacy reasons, we were unable to confirm if the outlier was a bot. An alternate (mathematical) explanation for this outlier is given in Appendix A. It is interesting to note that while the unconditional CCDF (Figure 1) has an inflection point near 100 friends, the conditional CCDF (Figure 5) has no such inflection point. This sheds light into the Dunbar number hypothesis applied to on-line social networks, first presented in [1] for the CyWorld network, which argues that a drastic drop in the CCDF near the Dunbar number 147.8 (such as the one in Figure 1) is a consequence of the theoretical cognitive limit of the number of people (about 147.8) with whom one can maintain stable social relationships [5]. As the Dunbar number should be valid for all "human" users, the hypothesis should clearly apply to the conditional CCDF as well. However, the conditional CCDFs (Figure 5) show no such point of inflection. On the other hand, the graph in Figure 7 shows these same conditional CCDFs (Figure 5) unconditioned upon the activity span (Figure 2). We see that the unconditional graph presents an inflection point near 130 friends. This indicates that the inflection point in Figure 1 is a consequence of the exponential activity spans and not a direct consequence of the Dunbar number.

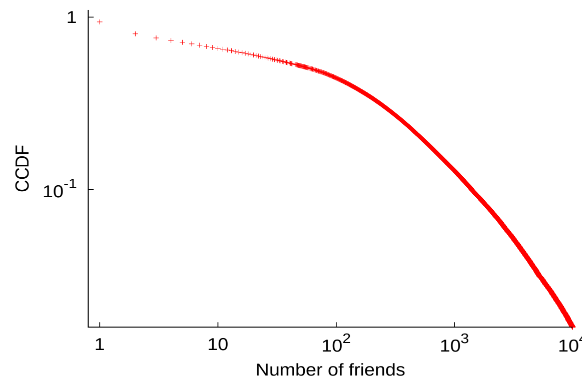


Fig. 7. Log-log plot of the CCDF of friends for accounts with activity spans (greater than 2 months) unconditioned on their account activity spans (also in months).

In order to test if the CCDF of activity spans from 3 to 65 months follow lognormal distribution we use the QQ-plots shown in Figure 6. Figure 6(a) shows all 63 QQ-plot curves from 3 months to 65 months. The straight line represents a perfect match to a lognormal distribution. Both axes in Figure 6 are not the number of friends as seen in a regular QQ-plot. If plotted in their regular scale, the regular QQ-curves cannot be compared in the same plot as they have different averages. We need to get around the fact that accounts with different activity spans have different lognormal parameters. Thus, we 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 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. Because many curves in Figure 6(a) intersect, we opt to also show, in Figure 6(b), the heatmap of Figure 6(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 distribution.

Estimates of the lognormal parameters (μ, σ) for each activity span value (in months) are plotted in Figure 8. Note that for activity spans greater than 4, the average number of friends μ seems to grow according to $\sqrt{T}$ while the standard deviation σ remains constant. In what follows we combine the above results to understand the distribution of friends in MySpace.

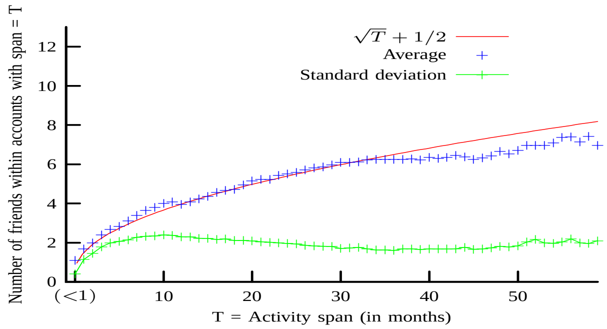


Fig. 8. The average number of friends as a function of the account activity span T grows according to $\sqrt{T}$ while the standard deviation remains constant.

IV. Double Pareto and the distribution of friends

Reed [11] shows that lognormally distributed random variables with parameters $(\mu T, \sigma^2 T)$ given a fixed activity span T, where T itself is a exponentially distributed random variable, results in a distribution that is characterized by its graph log-log CCDF plot: Two straight line segments that meet at a transition point [10]. Known as the Double Pareto distribution, we can see the connection between Reed's Double Pareto distribution and the distribution of MySpace friends shown in Figure 1.

The Double Pareto-like shape of MySpace' friends distribution is not surprising according to the statistics seen in Section III. In Section III we have shown that the distribution of MySpace friends shown in Figure 1 is a mixture of:

1) Exponential activity spans (Figure 2) (where active users log into their accounts quite frequently, as seen in Figure 4) with
2) the lognormal distribution of friends given a fixed activity span (Figure 6) with parameters $(\mu\sqrt{T}, \sigma^2)$ (Figure 8).

The parameters of the lognormals are the main difference between the statistics presented in this section and Reed's model [11]. In Reed's model the lognormal parameters are $(\mu T, \sigma^2 T)$ whereas in MySpace they are just $(\mu\sqrt{T}, \sigma)$. While this is not a trivial difference, in our case the end result is surprisingly similar, which deserves future study.

A. Lognormals and the Discrete Geometric Brownian Motion

First proposed by Huberman and Adamic [7] to explain the degree distribution of the WWW graph, later formalized by Reed [11] in the context of income distributions, and further extended to explain file size distributions [10], the model behind the Double Pareto distribution is a powerful tool that is able to connect power law distributions with Geometric Brownian Motion (GBM) processes. The following is a thought experiment based on the Huberman and Adamic [7] and Mitzenmacher [10] models that connects our results with a "discrete GBM model". In this model the standard deviation grows linearly with the account activity span, which seems to contradict our the empirical data, shown in Figure 8. We leave as future work understanding the odd behavior of the lognormal parameters in MySpace. Let X_d be a random variable that denotes the number of friends of a randomly chosen account with activity span of d days. Let

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

where F_i, $i = 1, 2, \ldots$ are independent random variables with finite mean and variance and $X_0 = 1$ (MySpace accounts 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[2] in equation (2), converges to the Normal distribution [14]. A direct consequence of the CLT is that the Normal distribution 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]. Thus, 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 explanation behind the lognormals seen in the conditional distributions of friends in Section III-B, although it does not explain the lognormal parameters seen in Figure 8. *It is important to note that this model makes no assumption on how friends connect to each other.*

V. Related work

Closely related to the above observation is the observation of Huberman and Adamic [7], 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 [7] 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 assuming 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 growth is replaced by the assumption of exponential activity spans (during which MySpace users are able to include and remove friends). Mitzenmacher [10] 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. The migration of MySpace users has been studied in [15] where the authors

[2] 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.

looked at the last login times. Our work, on the other hand, uses another metric (the activity span) to understand the distribution of the number of friends.

VI. Summary & Conclusions

In this work we studied the activity span of MySpace accounts and its connection to the distribution of the number of friends. We observed exponentially distributed activity spans and that the distribution of friends over accounts with the same activity spans can be well approximated by a lognormal with a fairly light tail. These new findings shed light into the puzzling (yet unexplained) shape of the distribution of friends in MySpace when plotted in log-log scale: Two Pareto-shaped distributions with different parameters joined by an inflection point (knee). We argued that the inflection point is more likely related to the inflection point of Reed's (Double Pareto) Geometric Brownian Motion with Exponential Stopping Times model than to the Dunbar number hypothesis of online social networks in [1]. While our work answers many question, we leave many others open, such as reason behind the puzzling constant standard deviation of the friends distribution conditioned on an activity span. Another related open question is if Reed's Geometric Brownian Motion with exponential stopping times model can be changed to accommodate lognormal distributions with parameters $(\mu\sqrt{T}, \sigma)$.

Appendix

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 [6]. In the complex networks literature heavy-tailed distributions are often found in observed (incomplete) data: links in Web pages [2], [7], file sizes [10], 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.

A. 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 [9]. 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 [6], 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.

In what follows we provide a detailed analysis of the estimation error and the maximum likelihood estimate of our data.

B. Estimation error

Let θ_i be the fraction of MySpace accounts with i friends and $\boldsymbol{\theta} = \{\theta_i | i = 1, \dots\}$. 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 [8]). 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

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

The above inequality is a straightforward application of the Cramér-Rao inequality [3]. 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 $d = 1, \dots$, 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. [6] 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.

C. Maximum likelihood estimation

Here we find 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

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}} P[\mathbf{Y} = \mathbf{y} \,|\, \boldsymbol{\theta}].$$

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 1 shows the empirical distribution of friends in our MySpace traces. The rather trivial conclusion is that the empirical curve seen in Figure 1 is the most likely distribution of friends in MySpace.

D. 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.

APPENDIX 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

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

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

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

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

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

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 $\sum_{\forall j} P[X = j] = 1$.

REFERENCES

[1] Yong-Yeol Ahn, Seungyeop Han, Haewoon Kwak, Sue Moon, and Hawoong Jeong. Analysis of topological characteristics of huge online social networking services. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 835–844, New York, NY, USA, 2007. ACM.

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

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

[4] 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.

[5] R. Dunbar. Coevolution of neocortex size, group size and language in humans. *Behavioral and Brain Sciences*, 16(4):681–735, 1993.

[6] 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.

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

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

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

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

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

[12] Sheldon M. Ross. The inspection paradox. *Probability in the Engineering and Informational Sciences*, 17(01):47–51, 2003.

[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.

[15] Mojtaba Torkjazi, Reza Rejaie, and Walter Willinger. Hot today, gone tomorrow: on the migration of MySpace users. In *WOSN '09: Proceedings of the 2nd ACM workshop on Online social networks*, pages 43–48, New York, NY, USA, 2009. ACM.