Are human interactivity times lognormal?

Norbert Blenn*and Piet Van Mieghem

Delft University of Technology
16 December 2015

Abstract

In this paper, we are analyzing the interactivity time, defined as the duration between two consecutive tasks such as sending emails, collecting friends and followers and writing comments in online social networks (OSNs). The distributions of these times are heavy tailed and often described by a power-law distribution. However, power-law distributions usually only fit the heavy tail of empirical data and ignore the information in the smaller value range. Here, we argue that the durations between writing emails or comments, adding friends and receiving followers are likely to follow a lognormal distribution.

We discuss the similarities between power-law and lognormal distributions, show that binning of data can deform a lognormal to a power-law distribution and propose an explanation for the appearance of lognormal interactivity times. The historical debate of similarities between lognormal and power-law distributions is reviewed by illustrating the resemblance of measurements in this paper with the historical problem of income and city size distributions.

1 Introduction

Massive data from online social media and online social networking (OSN) enables accurate analysis of the human behavior and of the interaction times between individuals through technology such as in word-of-mouth marketing, opportunistic networking and viral spreading. Empirical measurements contradict the commonly assumed exponential distribution of Poissonian inter-event times [1] in spreading processes or epidemic models. Multiple publications [2, 3, 4, 5, 6, 7, 8] report that *the interactivity time*, defined as

---

*Faculty of Electrical Engineering, Mathematics and Computer Science, P.O Box 5031, 2600 GA Delft, The Netherlands; *email*: N.Blenn@, P.F.A.VanMieghem@tudelft.nl

the duration between two consecutive tasks like sending emails, accessing web pages, instant messaging and phone calls, follow power-law distributions. These findings recently led to non-Markovian analyses, addressed in the work of Cator *et al.* [9], Iribarren and Moro [10], Van Mieghem and van de Bovenkamp [11] and Schweizer *et al.* [12]. Given that the inter-activity distributions are heavy tailed, word-of-mouth spreading, viral infections or dynamics of memes are expected to endure or survive longer, compared to basic Markovian models [10, 13]. Barabasi [2] infers that heavy-tailed distributions may arise from a priority queue, where individuals execute tasks of which the majority can be completed in short time, but some tasks wait long due to a perceived priority. Barabasi's priority queue model fits the distribution of durations between events quite well, leading to a power-law distribution with an exponent γ around 1.

A power-law random variable $X \geq \tau$ has the probability density function

$$f_X(t) = ct^{-\gamma} \qquad\qquad t \geq \tau \qquad\qquad (1)$$

where $c = \frac{1-\gamma}{\tau^{1-\gamma}}$ and $\tau > 0$ is the lower bound for X. The probability density function (pdf) of a lognormal random variable X for $t \geq 0$ is

$$f_X(t) = \frac{\exp\left[-\frac{(\log t - \mu)^2}{2\sigma^2}\right]}{\sigma t \sqrt{2\pi}} \qquad\qquad (2)$$

where (μ, σ) are called the parameters of the lognormal pdf, that are the mean and variance of $\log X$ as shown in Appendix A.

When we assume that inter-event durations are power-law distributed, we encounter the following issues:

1. In many cases, only a part of the data (the tail larger than τ) is modeled by a power-law. The lower bound τ in (1) does not correspond to the physical minimum of the random variable X, but τ is fitted from the data by ignoring smaller values that do not obey the power-law. Often, these smaller values may have a large probability to occur, so that their neglect is difficult to justify. In other words, only a part of the process (above τ) is modeled by a power-law (1), while the other part (below τ) is not.

2. Most processes or measurements possess both a lower as well as an upper bound. Apart from the lower bound τ, an upper bound κ is often invoked, at which a cut-off is observed: the power-law behavior is confined to the range $\tau \leq X \leq \kappa$, although $X_{\min} < \tau$ and $X_{\max} > \kappa$.

However, it is often unclear whether the process in the deep tail still obeys a power-law distribution or some other, much faster decreasing distribution. The upper bound κ is usually empirically determined, rather than based on the physical maximum of X. As long as

$$\Pr[X > \kappa] = c \int_{\kappa}^{\infty} t^{-\gamma} dt = \left(\frac{\kappa}{\tau}\right)^{1-\gamma}$$

is small (with respect to the measurement precision), the upper bound κ is justified, else other validation arguments are needed.

3. We demonstrate here that the binning of data (either by the data-provider or by the researcher) alters the shape of a lognormal distribution into an apparent power-law.

Particularly in relation to human activities or behaviors, we question in this paper the widely assumed power-law distribution.

We present measurements of inter-event durations from Digg.com and Reddit.com, two online social news aggregators [14, 15, 16], and from the Enron data set[1], a collection of emails sent by employees of the company Enron, and argue that a lognormal distribution is a valid candidate for the distribution of human inter-event durations in Section 2. The problem of fitting a lognormal is explained in Section 3, followed by previously reported lognormally distributed data sets in Section 4. Existing theoretical models are compared in Section 5, a plausible interpretation for lognormal human behavior is proposed in Section 6 and Section 7 concludes. Mathematical properties of the lognormal distribution are deferred to the Appendix A. Appendix B presents results of likelihood tests for the observed distributions.

2 Observations and Measurements from OSN

All events in an OSN are based on users' activities, such as posts, friend-requests and comments. A complete data set including all activities of users from Digg.com for a duration of 4 years, described in Tang *et al.* [14] and Doerr *et al.* [15], allows us to analyze the time frame in which users of Digg.com add their friends. Doerr *et al.* [19] found that reaction times in a retweet network from Twitter and Digg are close to a lognormal distribution with parameters $\mu = 10.1$ and $\sigma = 2.2$. The process of adding friends shows bursts of activity also observed and analyzed in email communication

---

[1]Enron Email Data-set, Leslie Kaelbling and Melinda Gervasio, http://www.cs.cmu.edu/∼enron/

[2, 10, 23, 17, 18]. In these publications the question arises, whether the observed distribution of inter-activity times is described by a power-law, lognormal or a cascading Poisson process.

Malmgren *et al.* [17, 18] describe that circadian and weekly activity cycles in human behavior are the factors that lead to heavy tails. They proposed a cascading Poisson process, consisting of a *nonhomogeneous* Poisson process that reflects periodicity and a *homogeneous* Poisson process describing active intervals, which is shown to model the interactivity distributions of e-mail communication.

Because the network of Digg.com is directed (like in Twitter.com or other OSNs), a user can be followed by other users to become their "friend", while a user cannot add followers. This means that the process of adding friends is solely based on the user himself, whereas obtaining followers depends on the activities of other users. The random variable T_{friend} denotes the time between the addition of two friends and, similarly, $T_{follower}$ is the time between receiving two followers.
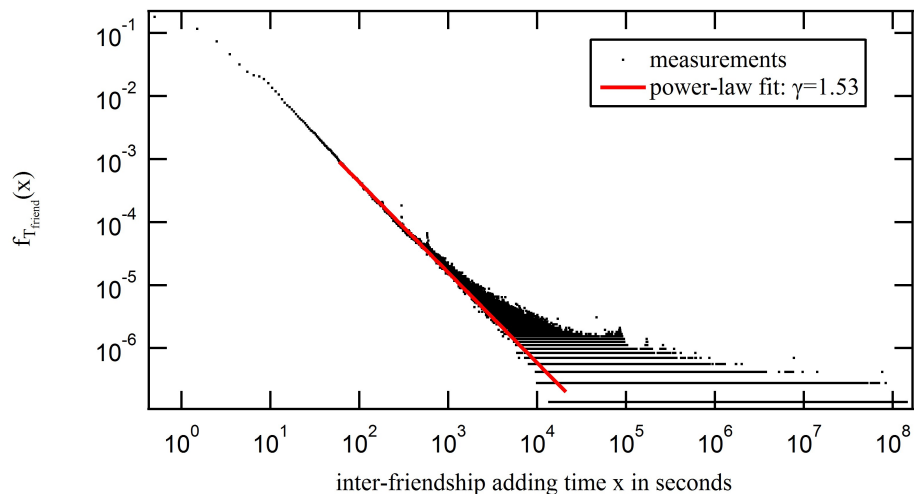

Figure 1: Time difference (in seconds) between the addition of two consecutive friends.

Figures 1 and 2 depict the distribution of durations between adding friends T_{friend} and of receiving followers $T_{follower}$ for 7.4 million friendship relations in Digg. Fitting the data, i.e. all realizations of T_{friend}, by the state of the art technique by Clauset *et al.* [20] to a power-law (1) results

4

in an exponent $\gamma = 1.53$ for $\tau = 59s$, whereas all realizations of $T_{follower}$ are fitted best by a lognormal (2) with parameters $\mu = 10.45$ and $\sigma = 2.75$. For $T_{follower}$ an estimated p-value of 0.0 indicates that a power-law fit provides not the best solution, whereas the p-value for T_{friend} of 0.23 indicates a reasonable fit for a power-law distribution.
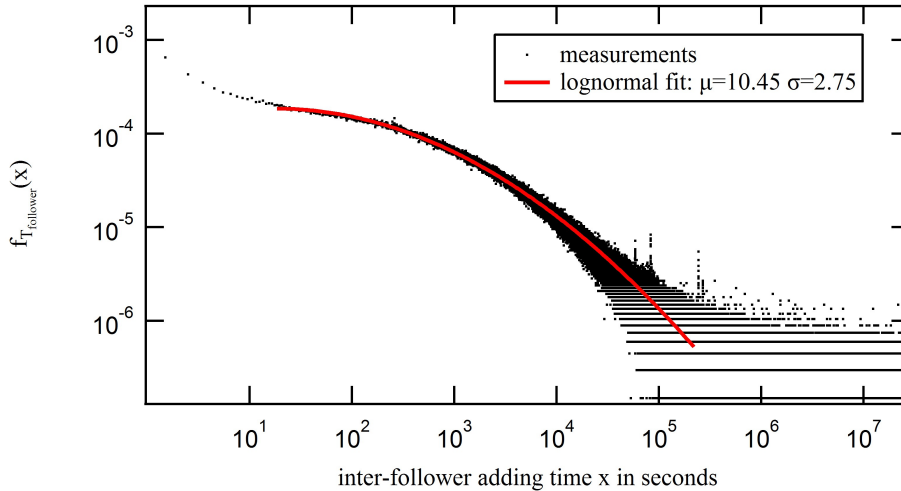

Figure 2: Time differences (in seconds) between the receipt of two consecutive followers.

The reason that the distribution of $T_{follower}$ does not fit the lognormal distribution over the whole range lies in the nature of Digg.com. Tang *et al.* [14] showed that only a few users were active over a long period, while just a fraction of them was actually submitting stories. Only 2% of all registered users succeeded to have their submissions "promoted" to the frontpage [15]. Since the username of a submitter appears next to the story, these users receive a lot of followers during the relatively small period that their story was listed on the front page. Therefore, $f_{T_{follower}}(x)$ is large for small x in Fig. 2. In summary, even ignoring the small time values, the random variable T_{friend} and $T_{follower}$ possess different distributions. The process that generates T_{friend} is more likely described by a variant of Barabasi's priority queue model, leading to power-law behavior. Since the random variable $T_{follower}$ is generated by the collective dynamics of different individuals (that are likely weakly dependent), central limit law arguments may point towards the lognormal distribution [1, p. 121-126].

5

Similar properties occur in sending and receiving emails. Obviously, a user can only send an email when he is online, whereas emails arrive at a user's inbox at his email server at all times. We analyzed the durations between receiving and sending emails in the Enron data set, which contains emails of all employees of Enron during roughly 6 years, starting in January 1998 until February 2004. Similar distributions arise as shown later in Figs. 9 and 10.

A third data set from Reddit.com[2], an OSN in which users mainly submit, comment or vote on bookmarks. The pdf of the duration T between consecutive comments and submissions in Reddit are shown in Fig. 3.
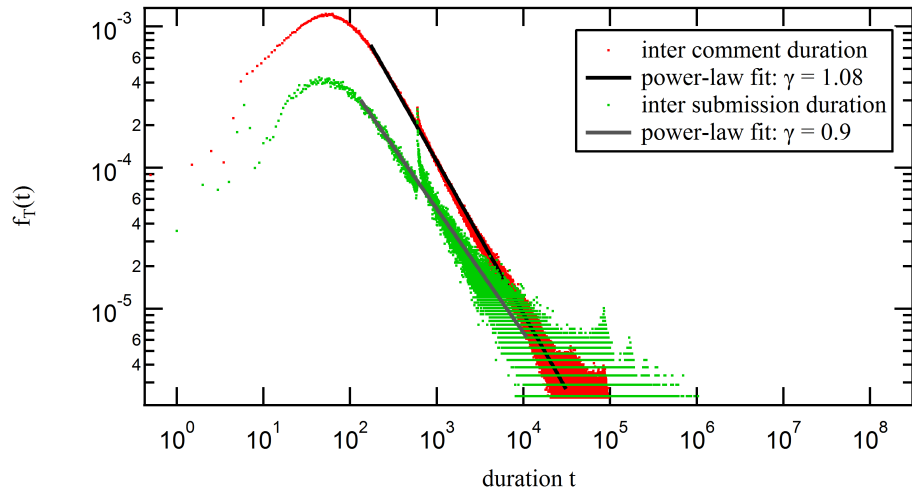

Figure 3: Time difference between commenting and bookmark submission in Reddit.com

The tails of the distributions in Fig. 3 seem nicely fitted by a power-law distribution with exponents $\gamma \approx 1$. However, as mentioned in the introduction, the increasing regime in $f_T(t)$ for small values of t, the peak of $f_T(t)$ nor the concave form of the pdf can be modeled by a power-law distribution.

---

[2]The Reddit.com data-set is hosted at Google BigQuery (bigquery.cloud.google.com/dataset/fh-bigquery).

3   Fitting a lognormal distribution

The two main approaches are based on the pdf and the EDF (empirical distribution function[3]), after a logarithmic transformation of the data. Figure 4 depicts the data fitted[4] to the distribution function of a normal distribution (4) and shows that the parameters of the EDF are about the same as those in Fig. 2 (pdf approach).
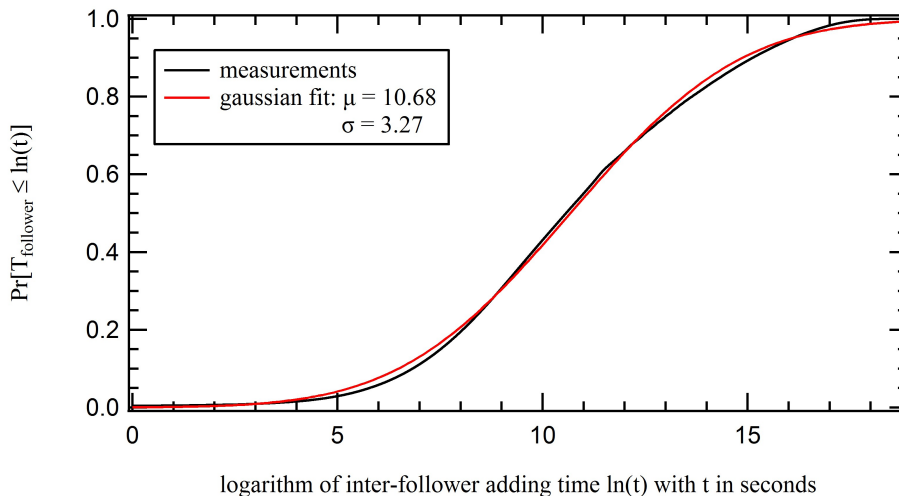

Figure 4: Empirical distribution function (EDF) of the inter-follower durations $T_{follower}$ (same data as in Fig. 2).

3.1   The effect of binning

3.1.1   Theory

We generate [1, p. 41] n realizations of a lognormal random variable X, which we denote by the set $\{x_k\}_{1 \leq k \leq n} = \{x_1, x_2, \ldots, x_n\}$. All n realizations lie in the interval $[x_{\min}, x_{\max}]$, where the minimum value is $x_{\min} = \min_{1 \leq k \leq n} x_k$ and the maximum value is $x_{\max} = \max_{1 \leq k \leq n} x_k$. The binning operation [1, p. 580-581] divides the entire data interval $[x_{\min}, x_{\max}]$

---

[3]The empirical distribution function is sometimes also called *empirical cdf*: the cumulative distribution function (cdf) based on an empirical measure.

[4]Fitting data to a EDF usually flattens interesting parts of a pdf, especially the tail of a distribution. Still, the benefit lies in the fact that binning is not needed and "raw" data can be directly fitted [1, p. 580-581].

into m sub-intervals of length $\Delta x = \frac{x_{\max} - x_{\min}}{m}$ and the j-th subinterval $[x_{\min} + (j-1)\,\Delta x, x_{\min} + j\Delta x]$ for $1 \leq j \leq m$ is associated with a bin h_j, that contains the number of realizations of X or data points of set $\{x_k\}_{1 \leq k \leq n}$ lying within the j-th subinterval,

$$h_j \approx n \int_{x_{\min}+(j-1)\Delta x}^{x_{\min}+j\Delta x} f_X(u)\, du$$

Clearly, for a given set $\{x_k\}_{1 \leq k \leq n}$, increasing the binsize Δx decreases the number of bins m.

The effect of binning on the pdf $f_X(t)$ is depicted in Fig. 5, where $n = 10^6$ realizations from a lognormal random variable with parameters $\mu = 10$ and $\sigma = 2$ are drawn. As demonstrated in Appendix A, the scaled random variable $Y = bX$ has parameter $\mu_Y = \mu_X + \ln b$, but $\sigma_Y = \sigma_X$, implying that "binning" only changes the parameter μ, but leaves the parameter σ invariant! By binning (scaling) the distribution with different binsizes, the "up-going regime", where $f_X(t)$ increases with t, disappears and the observable part of the distribution "evolves" towards a straight line on a log-log plot. Binning the data with larger binsizes decreases the parameter μ, even to the extent that μ may become negative. If $\mu - \sigma^2$ decreases, the maximum of the lognormal at $e^{\mu-\sigma^2}$ tends to zero (infinitely far to the left on a log-log scale). Therefore, just the decreasing part of the quadratic shape of (2) will be visible in a log-log plot.

3.1.2 Data analysis

Binning the data in different time units, say per minute instead of per second, scales the distribution by a factor of 60 (since 60 seconds equals 1 minute), which will shift the parameter $\mu_{\min}$ of the CDF towards the left to $\mu_{\min} = \mu_{\sec} - \ln(60) \approx \mu_{\sec} - 4.1$. However, a remarkable property of the lognormal distribution is that the parameter σ will not change after linear scaling as shown in the Appendix A.

An important consequence of a binning operation is illustrated in the pdfs in Fig. 6 and Fig. 7, that show the pdf of $T_{follower}$ binned per minute and per hour, respectively. Conforming to the theory, the parameter μ decreases by a factor of about $\ln 60 \approx 4.1$, while the parameter σ keeps its value, but the shape of the distribution changes. The distribution, binned per hour, is shown in Fig. 7, where the distribution can be confounded with a power-law distribution (with exponential cut-off). Moreover, this misinterpretation may be justified, because the exponent $\gamma = 1.57$ of a power-law fitted to the hourly binned data is close to the exponent $\gamma = 1.53$
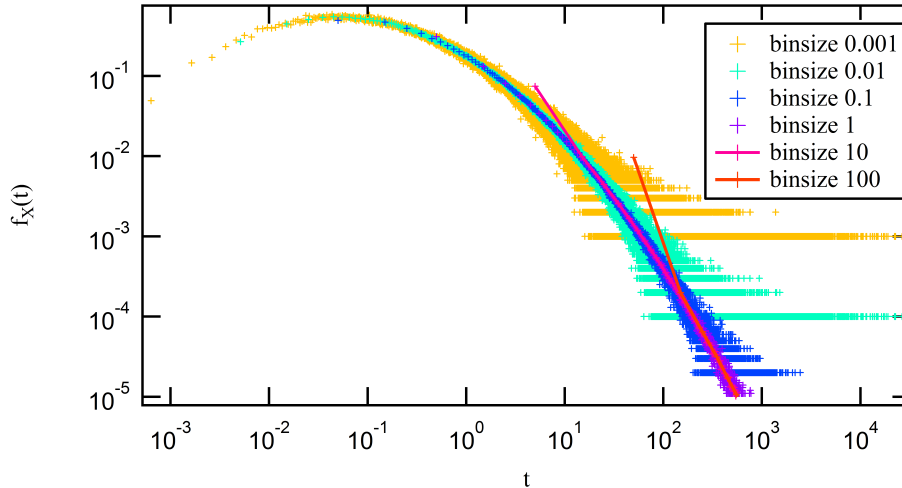
Figure 5: The effect of different binsizes on a lognormal random variable

found for the T_{friend} distribution in Fig. 1. Fig. 8 shows the complementary cumulative distribution function of $T_{follower}$ binned per hour, fitted using the method of Clauset *et al.* [20]. The resulting power-law exponent in Fig. 8 is $\gamma = 1.81$. The visualization in Fig. 8 is misleading because the original distribution is, as shown earlier, a lognormal distribution.

For completeness we like to refer to Table 3 in Appendix B, where we list fitted parameters, p values and the result of log-likelihood tests, conducted using the method of Virkar and Clauset [51]. As shown in the table, the larger the used binsizes, the higher the p values indicating that a power-law fits the data. For small binsizes (low p values) the results are not conclusive or rather in favor of a lognormal distribution.

To further demonstrate the effect, Figs. 6, 7 and 8 depict distributions in which the data was artificially binned by minutes and hours. In reality, such binning operation might occur if the data provider (the OSN or web service) returns values per hour or even larger scales. In other words, data measured in larger time steps over a certain total time range thus erases the possibility to distinguish between a lognormal and power-law distribution.

The sending time of an email in the Enron data set until May 2001 was stored per minute, whereas afterwards (June 2001 - February 2004) the sending time is stored with an accuracy of a second. The resulting pdf of time differences between sent emails is plotted in Fig. 9, where the
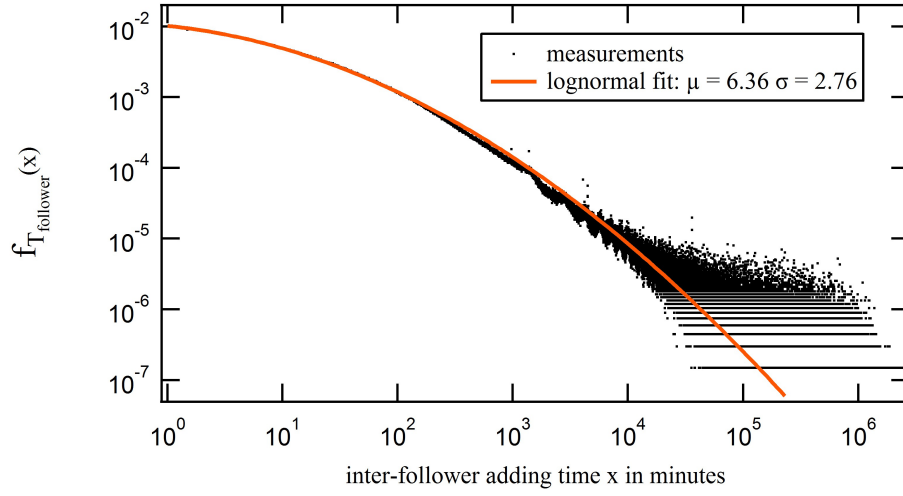
Figure 6: The probability density function (pdf) of the inter follower times $T_{follower}$ binned per minute.
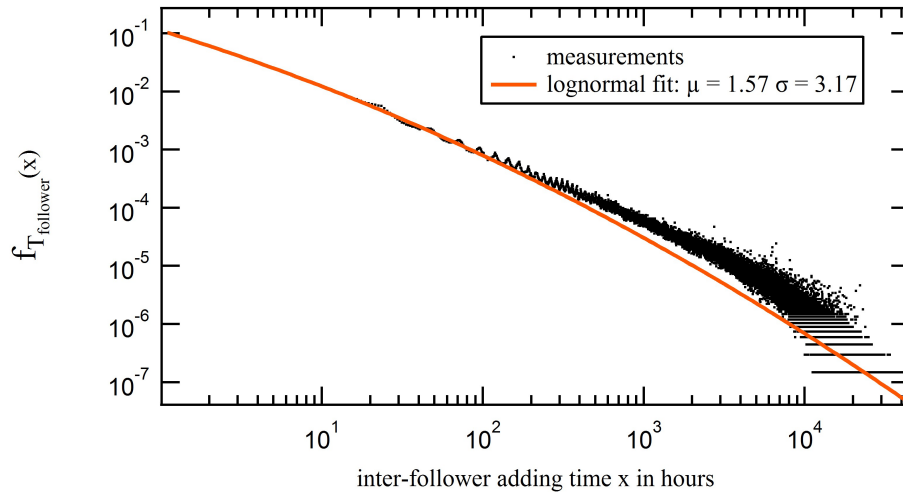

Figure 7: The probability density function (pdf) of the inter follower times $T_{follower}$ binned per hour.
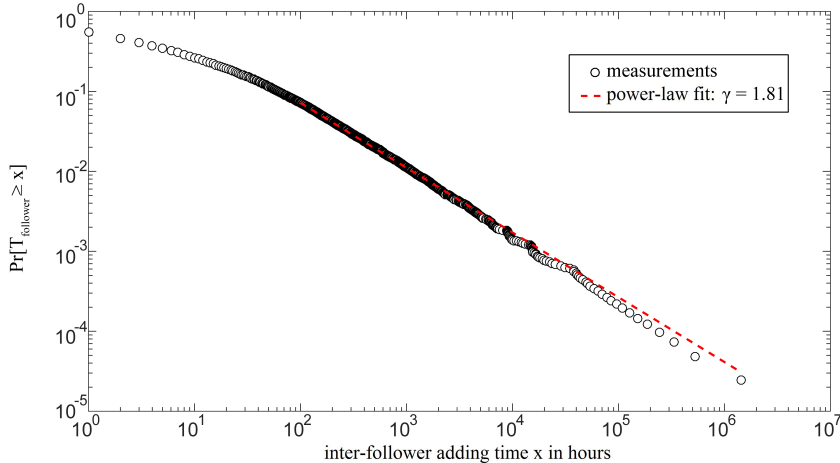
10

Figure 8: The ccdf of the inter-follower time $T_{Follower}$ where data is binned per hour.

black and red dots represent the measurements per second and per minute, respectively. The parameters of the fitted lognormal ($\mu = 9.5, \sigma = 3.24$) and power-law distribution ($\gamma \approx 1$) correspond to those of Digg.com in Fig. 2 and Fig. 1.

The pdf of inter-arrival times of received emails, split in observations per minute (red) and per second (black) in Fig. 10, can be compared to the durations of added followers, where the parameters are again in the same range as shown before. We found that the EDF of the data (not shown) agrees with the pdfs in Fig. 9 and Fig. 10, underlining the likeliness that a lognormal random variable models the data better than a power-law.

4   Reported parameters of power-law and lognormal distributions

Table 1 lists inter-arrival times and the fitted exponents found in publications in which the aspect of binning was ignored. Rather small power-law exponents γ between 0.7 and 1.8 are found in multiple data sets of human activity.

Lognormal distributions were reported for similar human activity as shown in Table 2, which extends the collection of Limpert *et al.* [21]. The
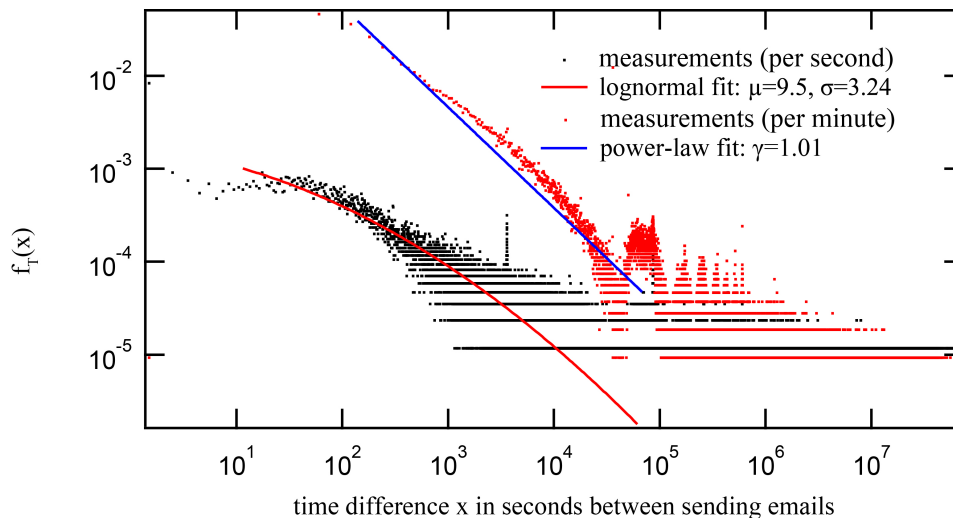
Figure 9: Time difference between consecutively sent emails per person in the Enron data set.

oldest analysis was conducted by Boag [22] in 1949. Based on the scaling invariance of σ as demonstrated in Appendix A, the parameter σ can be compared over different measurements, whereas the parameter μ cannot, since μ depends on the units in which the lognormal random variable is measured. Interestingly, all σ's in Table 2 lie within an amazingly small range of $0.35 \leq \sigma \leq 3.2$, which shows that the parameter σ only varies over about one order of magnitude in different measured phenomena. If σ is rather small as in our measurements where $2.73 \leq \sigma \leq 3.24$, the first term in (10) dominates and a quadratic function appears in a log-log plot. Consequently, these rather small values of σ and those in Table 2 contradict the common deductions made from (10) in Appendix A, namely that only for large values of σ power-law and lognormal distributions are indistinguishable. Equations (8) and (9) in Appendix A explain why the lognormal distribution may seem linear in a certain regime.

5 The Historical Debate of the Power-law versus Lognormal Distribution

Mitchell [40] discusses the fascinating and abundant appearance of power-law distributions and mentions processes such as self-criticality, that are
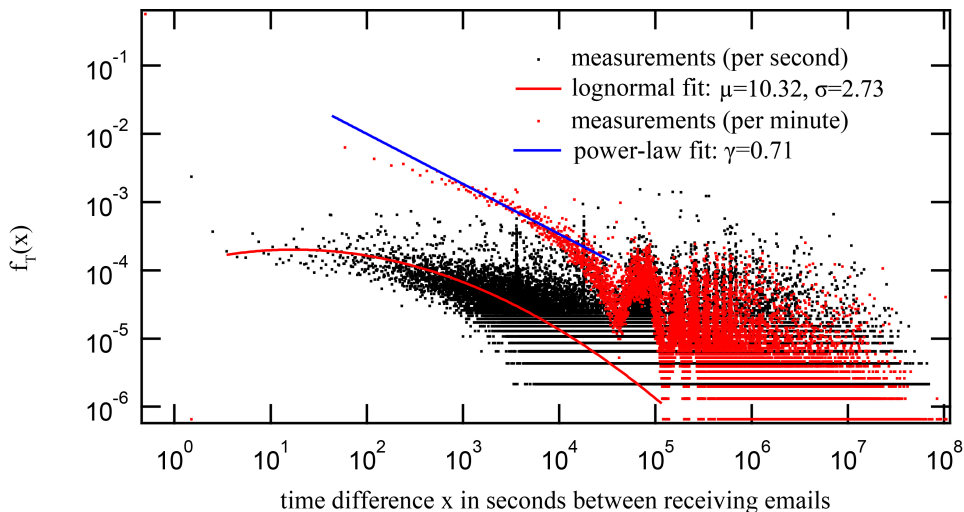
12

Figure 10: Time difference between consecutively received emails per person in the Enron data set.

related to phase transitions, as the main producers of power-laws, but she remarks that the precise nature as well as the deviations from a power-law are still open to debate. Table 2 illustrates that the research on lognormal distributions has a long history. Unfortunately, also the processes that produce (almost) lognormal behavior are not well understood. Mitzenmacher [41] overviews processes leading to power-law and lognormal distributions, emphasizing that minor changes in the process will lead to either a power-law or a lognormal distribution. Mitzenmacher mentions the work of Gabaix [42], who analyzed the size distribution of cities in the United States. Interestingly, Gabaix [42] found that the city size distribution follows a Zipf distribution, which is a power-law distribution (1) with an exponent $\gamma = 1$. Gabaix argues that cities cannot become infinitely small, a fact that imposes a lower bound to their size. When modeling the size of cities as a Markov chain with a fixed number of cities, which grow stochastically as proposed by Gibrat [43], then the steady state of the Markov chain will follow Zipf's distribution with an exponent $\gamma = 1$. If there is no lower bound on the city size, so that cities can be arbitrarily small, then the distribution degenerates to a lognormal distribution.

Gibrat [43], whose work is often associated with the law of proportionate effect [39], argues in his model that transition probabilities or the variance

13

Data-set	Exponent (γ)	Reference
Time between sending Emails	1	Barabasi (2005) [2]
Time between sending Emails	1.2	Eckmann *et al.* (2004) [3]
Time between clicks in website usage	1, 1.25	Gonçalves and Ramasco (2008) [4]
Time between similar actions in web data	1.1, 1.2, 1.8	Radicchi (2008) [5]
Time between messages in instant-messaging	1.53	Leskovec and Horvitz (2008) [6]
Time between phone calls	0.9	Candia *et al.* (2008) [7]
Time between phone calls	0.7	Karsai *et al.* (2012) [8]
Time between sending Emails	1	Karsai *et al.* (2012) [8]
Time between sending short messages	0.7	Karsai *et al.* (2012) [8]

Table 1: Power-law exponents found for durations between technology related human dynamics

of transition probabilities in a growth process are independent from the size. Gibrat, who estimated the distribution of city sizes, concluded that the law of proportionate growth leads to a distribution that is lognormal. However, Simon [44] showed that Gibrat's law of the proportionate effect may lead to other heavy-tailed distributions as well.

Champernowne [45] and Cordoba [46] analyzed the income distribution of England and Wales with Markov theory. Again, their crucial assumption is that incomes have a lower bound.

Eeckhout [47] analyzed the distribution of city sizes by using accurate data from the US census in 2000 and found that the heavy tail obeys Zipf's law. But, the entire distribution is better described by a lognormal than a Pareto distribution. By comparing the census data from 1990 and 2000, Eeckhout shows that the growth of a city is independent of its size. The parameters of the lognormal distribution found by Eeckhout are $\mu = 7.28$ and $\sigma = 1.75$.

The difference with the above mentioned Markov chain approach lies in the fact that Eeckhout [47] modeled the process by a multiplicative process, which leads to a lognormal distribution. This multiplicative process, proposed by Kapteyn [48] in 1903 for the first time and later coined the "Law of Proportionate Effect" by Gibrat [43], is based on the central limit theorem applied to a multiplicative process, which leads to lognormal distributed sizes [1]. Gibrat's growth process is defined as

$$S_t = a_t \times S_{t-1} \qquad (3)$$

where the size S_t of an item at state t depends upon the previous size S_{t-1} times a positive, random factor a_t. By taking the logarithm of both sides in

Table 2: Literature of lognormal distributions (excerpt)

μ	σ	process	reference
5.547	2.126	email forwarding	Iribarren and Moro [10]
≈ 8	≈ 2	email forwarding	Stouffer *et al.* [23]
$\mu_1 = 1$ hour			
$\mu_2 = 2$ days		email forwarding	Stouffer *et al.* [24]
2.47	0.38	infection times	Nishiura [25]
2.47	0.36	latency periods of diseases	Limpert [21]
14 days	1.14	latency periods of diseases	Sartwell [26]
100 days	1.24	latency periods of diseases	Sartwell [26]
2.3 hours	1.48	latency periods of diseases	Sartwell [26]
2.4 days	1.47	latency periods of diseases	Sartwell [26]
12.6 days	1.50	latency periods of diseases	Sartwell [26]
21.4 days	2.11	latency periods of diseases	Sartwell [26]
9.6 months	2.5	survival times after cancer diagnosis	Boag [22]
15.9 monts	2.8	survival times after cancer diagnosis	Feinleib and Macmahon [27]
17.2 months	3.21	survival times after cancer diagnosis	Feinleib and Macmahon [27]
14.5 months	3.02	survival times after cancer diagnosis	Boag [22]
60 years	1.16	age of onset of Alzheimer	Horner [28]
4 days		incubation periods (viral infections)	Lessler *et al.* [29]
3 to 5	≈ 2	task completion	Linden [30]
0.5	1.4	strike duration	Lawrence [31]
		time of individual activities	Mohana *et al.* [32]
0.43	1.634	call duration	Spedalieri *et al.* [33]
3.5	0.70	message holding time	Barcelo and Jordán [34]
7.439	0.846	transmission holding time	Barcelo and Jordán [34]
3.29	0.890	channel holding time	Barcelo and Jordán [34]
3.3	0.89	channel holding time	Barcelo and Jordán [34]
$\mu_1 = 1.31$	$\sigma_1 = 0.32$		
$\mu_2 = 2.26$	$\sigma_2 = 0.56$	call holding time	Bolotin [35]
		citations	Eom and Fortunato [36]
		citations	Redner [37]
1 to 2	$0.35 - 0.45$	citations	Stringer *et al.*[38]
	1.095	citations	Radicchi *et al.*[55]
$\mu_1 = 3.7$	$\sigma_1 = 0.8$		
$\mu_2 = 5.6$	$\sigma_2 = 3.1$	retweeting behavior	Doerr *et al.*[19]
5.29	0.42	distribution of votes on pages of Digg.com	Van Mieghem *et al.* [39]

(3) and denoting $\xi_t \equiv \ln a_t$, we obtain, after iteration,

$$\ln S_t = \ln S_{0,i} + \xi_1 + \xi_2 + \ldots + \xi_t$$

By the Central Limit Theorem [1], $\frac{\sum_{k=1}^{t} \xi_k - t\mu}{\sigma\sqrt{t}} \xrightarrow{d} N(0,1)$, we arrive at, for large t,

$$\Pr[\ln S_t \leq y] \rightarrow \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{y-t\mu}{\sigma\sqrt{t}}} e^{-u^2/2} \, du$$

from which approximately

$$\ln S_t \cong t\mu + \sqrt{t}\sigma N(0,1)$$

where μ denotes the mean and σ^2 the variance of the sequence $\{\xi_k\}_{1 \leq k \leq t}$.

For large t, a power-law distribution (1) tends to zero as $O\left(t^{-\gamma}\right)$, while a lognormal distribution (2) tends considerably faster as $O\left(t^{-1}\exp\left[-\frac{\log^2 t}{2\sigma^2}\right]\right)$ to zero, illustrating that the deep tails are significantly different. Malevergne [49] addressed this fact by the concept of a slowly varying function, in particular, for $x \to \infty$ and $t > 0$, the power-law distribution (1) features

$$\lim_{x \to \infty} \frac{f_X(t \cdot x)}{f_X(x)} = t^{-\gamma}$$

The lognormal distribution (2) on the other hand is not slowly varying. In the limit $x \to \infty$, a lognormal distribution will, for $t > 1$, always tend to zero:

$$\lim_{x \to \infty} \frac{f_X(t \cdot x)}{f_X(x)} = \lim_{x \to \infty} \frac{1}{t} e^{-\frac{(\ln(t))^2}{2\sigma^2}} e^{-\ln(t) \cdot \frac{\ln(x)-\mu}{\sigma^2}} = 0$$

This different tail behavior questions whether exponential bin sizes should be used in a log-log plot, because these may hide the rapid decrease in the tail.

6   An explanation of lognormal human interactivity times

Perhaps, a plausible explanation of the appearance of lognormal human interactivity times is that the logarithm $\log T$ of a human interactivity time T is Gaussian or normally distributed (see Appendix A), most likely as a consequence of the Central Limit Theorem. Roughly speaking, the Central Limit Theorem applies for a large number of weakly dependent random

variables, where none of them is dominant. Thus, rather than concentrating on the time T, it seems more natural to focus on the random variable $\log T$ as the decisive quantity. The logarithm of a human related measure often occurs: for example, in our hearing system, the intensity of sound is logarithmically experienced. But more importantly, the large variability in human performance also seems logarithmically distributed [50, 30] and the measured interactivity times are strongly related to the large differences in human performance and/or behavior.

On the other hand, if Y is an exponential random variable with mean $\frac{1}{\gamma-1}$ for $Y \geq \log \tau$ (else $Y = 0$) and $X = e^Y$, then

$$\Pr\left[X \leq t\right] = \Pr\left[Y \leq \log t\right] = 1 - e^{-(\gamma-1)(\log t - \log \tau)} = 1 - \left(\frac{t}{\tau}\right)^{1-\gamma}$$

so that $\frac{d}{dt}\Pr\left[X \leq t\right] = \frac{\gamma-1}{\tau^{1-\gamma}}t^{-\gamma}$ is a perfect power-law probability density function as in (1). However, if $\log T \geq \log \tau$ were exponentially distributed with mean $\frac{1}{\gamma-1}$, the memoryless property of the exponential distribution [1, p. 43] would indicate that

$$\Pr\left[\log T \geq t + u \mid \log T > u\right] = \Pr\left[\log T \geq t\right]$$

In other words, given that the logarithm of a human interactivity time is larger than u time units, the probability that $\log T$ exceeds $t+u$ time units is equal to the probability that $\log T$ exceeds t time units, for any u and t larger than $\log \tau$, precisely as if the log-interactivity time u never has been spent or waited, which is quite counterintuitive for a duration between consecutive activities. Alternatively, with $u = \log s$ and $s > \tau$, the memoryless property of $\log T$ implies "scale-freeness in T":

$$\Pr\left[\log \frac{T}{s} \geq t \mid \log \frac{T}{s} > 0\right] = \Pr\left[\log T \geq t\right]$$

which shows, ignoring the condition $\log \frac{T}{s} > 0$, the independence on the "scale s". But, we have shown in Section 3.1 that rescaling the human interactivity time (by different bin sizes) definitely alters the distribution.

From these arguments, we can infer that a power-law time T is less defendable than a lognormally distributed T.

7 Conclusion

In this paper, the interactivity durations of individuals, between creating friendship relations, writing emails, commenting and voting on online content are analyzed. We found that the distribution of durations to add friends

follows a power-law with an exponent of $\gamma \simeq 1.8$, whereas the durations to acquire followers are well described by a lognormal with $\mu \approx 10.5$ and $\sigma \approx 2.8$. Due to the small probability of executing two tasks in a small time interval (typically ignored in fitting a power law), we claim that a lognormal distribution covers the entire activity time range better than a power distribution.

In addition, we show that binning of lognormally distributed data can seriously affect the perception: the parameter μ shifts towards smaller values, but the parameter σ of a lognormal distribution does not change after a binning or scaling operation. In the extreme case, only the heavy tail of the lognormal distribution (2) remains, which follows a power-law distribution (1) with an exponent of γ close to 1. As explained in the Appendix, there exists an interval in which the lognormal distribution is indistinguishable from a power-law distributions with power-law exponent $\gamma \approx 1 + \varepsilon$, for small $\varepsilon > 0$.

Similar observations and concerns, discussed for a long time in the literature of city size and income distributions and reviewed in Section 5, supports that a lognormal distribution is modeling the whole data range better than power-laws. Finally, Section 6 argues that the logarithm of a quantity associated with human activities rather than the quantity itself is the better descriptor, because human performance seems to fit a lognormal.

References

[1] P. Van Mieghem. *Performance Analysis of Communications Systems and Networks*. Cambridge University Press, 2014.

[2] A.-L. Barabási. The origin of bursts and heavy tails in human dynamics. *Nature*, 435:207–211, May 2005.

[3] J.-P. Eckmann, E. Moses, and D. Sergi. Entropy of dialogues creates coherent structures in e-mail traffic. *Proceedings of the National Academy of Sciences of the United States of America*, 101(40):14333–14337, 2004.

[4] B. Gonçalves and J. J. Ramasco. Human dynamics revealed through Web analytics. *Physical Review E*, 78.2026123, 2008.

[5] F. Radicchi. Human activity in the web. *Physical Review E*, 80.026118, Aug 2009.

[6] J. Leskovec and E. Horvitz. Planetary-scale views on a large instant-messaging network. In *Proceeding of the 17th international conference on World Wide Web*, pages 915–924. ACM, 2008.

[7] J. Candia, M. C. González, P. Wang, T. Schoenharl, G. Madey, and A.-L. Barabási. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 41(22):224015, 2008.

[8] M. Karsai, K. Kaski, A.-L. Barabási, and J. Kertész. Universal features of correlated bursty behaviour. *Scientific Reports*, 2:397, May 2012.

[9] E. Cator, R. van de Bovenkamp, and P. Van Mieghem. Susceptible-infected-susceptible epidemics on networks with general infection and cure times. *Physical Review E*, 87:062816, Jun 2013.

[10] J. L. Iribarren and E. Moro. Impact of Human Activity Patterns on the Dynamics of Information Diffusion. *Physical Review Letters*, 103(3):038702, Jul 2009.

[11] P. Van Mieghem and R. van de Bovenkamp. Non-Markovian Infection Spread Dramatically Alters the Susceptible-Infected-Susceptible Epidemic Threshold in Networks. *Physical Review Letters*, 110:108701, Mar 2013.

[12] I. Scholtes, N. Wider, R. Pfitzner, A. Garas, C. J. Tessone, and F. Schweitzer. Causality-driven slow-down and speed-up of diffusion in non-markovian temporal networks. *Nature Communications*, 5:5024, September 2014.

[13] M. Karsai, M. Kivelä, R. K. Pan, K. Kaski, J. Kertész, A. L. Barabási, and J. Saramäki. Small but slow world: How network topology and burstiness slow down spreading. *Physical Review E*, 83(2):025102, Feb. 2011.

[14] S. Tang, N. Blenn, C. Doerr, and P. Van Mieghem. Digging in the Digg Social News Website. *IEEE Transactions on Multimedia*, 13(5):1163–1175, 2011.

[15] C. Doerr, N. Blenn, S. Tang, and P. Van Mieghem. Are Friends Overrated? A Study for the Social News Aggregator Digg.com. In *Computer Communications*, volume 35, pages 796–809, 2012.

[16] P. Van Mieghem. Human psychology of common appraisal: The reddit score. *IEEE Transactions on Multimedia*, 13(6):1404–1406, 2011.

[17] R. D. Malmgren, D. B. Stouffer, A. E. Motter and L. A. N. Amaral A Poissonian explanation for heavy tails in e-mail communication. *Proceedings of the National Academy of Sciences*, 105(47):18153, 2008.

[18] R. D. Malmgren, D. B. Stouffer, A. S. L. O. Campanharo and L. A. N. Amaral On universality in human correspondence activity. *Science*, 325(5948):1696–1700, 2009.

[19] C. Doerr, N. Blenn, and P. Van Mieghem. Lognormal Infection Times of Online Information Spread. *PLoS ONE*, 8(5):e64349, 05 2013.

[20] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-Law Distributions in Empirical Data. *SIAM review*, 51(4):661–703, Nov. 2009.

[21] E. Limpert, W. A. Stahel, and M. Abbt. Log-normal distributions across the sciences: Keys and clues. *BioScience*, 51(5):341–352, 2001.

[22] J. W. Boag. Maximum Likelihood Estimates of the Proportion of Patients Cured by Cancer Therapy. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11(1), 1949.

[23] D. B. Stouffer, R. D. Malmgren, and L. A. N. Amaral. Comment on The origin of bursts and heavy tails in human dynamics. arXiv:physics/0510216, 2005.

[24] D. B. Stouffer, R. D. Malmgren, and L. A. N. Amaral. Log-normal statistics in e-mail communication patterns, arXiv:physics/0605027 2006.

[25] H. Nishiura. Early efforts in modeling the incubation period of infectious diseases with an acute course of illness. *Emerging Themes in Epidemiology*, 2007.

[26] P. E. Sartwell. The distribution of incubation periods of infectious disease. *American Journal of Hygiene*, 51:310–318, 1950.

[27] M. Feinleib and B. Macmahon. Variation in the duration of survival of patients with the chronic leukemias. *Blood*, 15, 1960.

[28] R. D. Horner. Age at onset of Alzheimer's disease: clue to the relative importance of etiologic factors? *American journal of epidemiology*, 126(3):409–14, 1987.

[29] J. Lessler, N. G. Reich, R. Brookmeyer, T. M. Perl, K. E. Nelson, and D. A. Cummings. Incubation periods of acute respiratory viral infections: a systematic review. *The Lancet infectious diseases*, 9(5):291–300, May 2009.

[30] W. J. van der Linden. A Lognormal Model for Response Times on Test Items. *Journal of Educational and Behavioral Statistics*, 31(2), 2006.

[31] R. J. Lawrence. The Lognormal Distribution of the Duration of Strikes. *Journal of the Royal Statistical Society. Series A*, 147(3), 1984.

[32] S. Mohan, M. Gopalakrishnan, H. Balasubramanian, and A. Chandrashekar. A lognormal approximation of activity duration in PERT using two time estimates. *Journal of the Operational Research Society*, 58(6):827–831, 2006.

[33] A. Spedalieri, I. Martín-Escalona, and F. Barceló. Simulation of teletraffic variables in UMTS networks: impact of lognormal distributed call duration. In *Wireless Communications and Networking Conference*, pages 2381–2386. IEEE, 2005.

[34] F. Barceló and J. Jordán. Channel Holding Time Distribution In Cellular Telephony. In *Electronics Letters, Vol.34 No.2*, pages 146–147, 1998.

[35] V. A. Bolotin. Modeling Call Holding Time Distributions for CCS Network Design and Performance Analysis. *IEEE Journal on Selected Areas in Communications*, 12:433–438, 1994.

[36] Y.-H. Eom and S. Fortunato. Characterizing and Modeling Citation Dynamics. *PLoS ONE*, 6(9):e24926, Sept. 2011.

[37] S. Redner. Citation Statistics from 110 Years of Physical Review. *Physics Today*, 58(6):49–54, 2005.

[38] M. Stringer, M. Sales-Pardo, and L. A. N. Amaral. Effectiveness of journal ranking schemes as a tool for locating information. *PLoS One*, 3(2):e1683, 2008.

[39] P. Van Mieghem, N. Blenn, and C. Doerr. Lognormal distribution in the digg online social network. *The European Physical Journal B - Condensed Matter and Complex Systems*, 83:251–261, 2011. 10.1140/epjb/e2011-20124-0.

[40] M. Mitchell, *Complexity: A Guidede Tour*, Oxford University Press, 2009.

[41] M. Mitzenmacher. A Brief History of Generative Models for Power Law and Lognormal Distributions. *Internet Mathematics*, 1(2), 2003.

[42] X. Gabaix. Zipf's Law For Cities: An Explanation. *Quarterly Journal of Economics*, 114:114–739, 1999.

[43] R. Gibrat. *Les Inégalités économiques [The Economic Inequalities]*. Librairie du Recueil Sirey, 1931, French.

[44] H. A. Simon. On a class of skew distribution functions. *Biometrika*, 42(3–4):425–440, 1955.

[45] D. G. Champernowne. A Model of Income Distribution. *The Economic Journal*, 63(250):pp. 318–351, 1953.

[46] J. C. Cordoba. A Generalized Gibrat's Law. *International Economic Review*, 49(4):1463–1468, 2008.

[47] J. Eeckhout. Gibrat's Law for (All) Cities. *American Economic Review*, 94(5):1429–1451, September 2004.

[48] J. C. Kapteyn. Skew Frequency curves in Biology and Statistics. *Molecular and General Genetics MGG*, 19(3):205–206, 1918.

[49] Y. Malevergne, V. Pisarenko, and D. Sornette. Gibrats law for cities: uniformly most powerful unbiased test of the Pareto against the lognormal. Swiss Finance Institute Research Paper Series 09-40, Swiss Finance Institute, 2009.

[50] W. Shockley, "On the statistics of individual variations of productivity in research laboratories", Proceedings of the Institute of Radio Engineers (IRE), Vol. 45, No. 3, pp. 279-290, 1957.

[51] Y. Virkar and A. Clauset. Power-law distributions in binned empirical data. *The Annals of Applied Statistics*, 8(1):89–119, 03 2014.

[52] F. Black and M. S. Scholes. The Pricing of Options and Corporate Liabilities. *Journal of Political Economy*, 81(3):637–54, May-June 1973.

[53] D. Garcia, P. Mavrodiev, and F. Schweitzer. Social resilience in online communities: The autopsy of friendster. In *Proceedings of the first ACM conference on Online social networks*, ACM, February 2013.

[54] H. Kesten. Random Difference Equations and Renewal theory for product of Random Matrices. *Acta Mathematica*, CXXXI:207–248, 1973.

[55] F. Radicchi, S. Fortunato, and C. Castellano. Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences*, 105(45), 2008.

A The lognormal random variable and distribution

A lognormal random variable [1, p. 57] is defined as $X = e^Y$ where $Y = N(\mu, \sigma^2)$ is a Gaussian or normal random variable. Hence, $X \geq 0$. The distribution function $F_X(t) = \Pr[X \leq t] = \Pr[Y \leq \log t]$ is

$$F_X(t) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\log t} \exp\left[-\frac{(u-\mu)^2}{2\sigma^2}\right] du = \frac{1}{2}\left(1 + \text{erf}\left(\frac{t-\mu}{\sigma\sqrt{2}}\right)\right) \quad (4)$$

where $\text{erf}(x)$ is the error function. The probability density function (pdf) of a lognormal random variable X follows from the definition $f_X(t) = \frac{d}{dt}\Pr[X \leq t]$, for $t \geq 0$, as (2), where (μ, σ) are called the parameters of the lognormal pdf, while the mean and variance are [1, p. 57]

$$E[X] = e^\mu e^{\frac{\sigma^2}{2}}$$

and

$$\text{Var}[X] = e^{2\mu} e^{\sigma^2}\left(e^{\sigma^2} - 1\right)$$

The limit $\sigma \to 0$ reduces to a Dirac delta function at $t = e^\mu$, thus $\lim_{\sigma \to 0} f_X(t) = \delta(t - e^\mu)$.

Given the mean and variance, the parameters of the lognormal are found as

$$\sigma^2 = \log\left(1 + \frac{\text{Var}[X]}{(E[X])^2}\right) \quad (5)$$

and

$$\mu = \log E[X] - \frac{\sigma^2}{2} \quad (6)$$

Although $E[X] \geq 0$, we remark that the parameter μ can be negative. Moreover, (5) and (6) show that the scaled lognormal random variable $Y = bX$, where b is a positive real number, has mean $\sigma_Y = \sigma$ and $\mu_Y = \mu + \log b$. Hence, *scaling* by a factor b does not change the parameter σ, which

has interesting consequences for binning and measured data: the unit (e.g. second versus hours) in which the random variable is measured does not alter the parameter σ, only the parameter μ.

The change of the argument $t \to e^u$ in $f_X(t)$ leads to

$$f_X(e^u) = e^{-\mu + \frac{\sigma^2}{2}} \frac{\exp\left[-\frac{\left(u - \left(\mu - \sigma^2\right)\right)^2}{2\sigma^2}\right]}{\sigma\sqrt{2\pi}} \tag{7}$$

illustrating that the scaled lognormal pdf $e^{\mu - \frac{\sigma^2}{2}} f_X(e^u)$ is a Gaussian pdf $N\left(\mu', \sigma^2\right)$ with mean $\mu' = \mu - \sigma^2$. The maximum of $f_X(t)$ occurs at $t = e^{\mu - \sigma^2}$ and equals $\max_{t \geq 0} f_X(t) = \frac{e^{-\mu} e^{\frac{\sigma^2}{2}}}{\sigma\sqrt{2\pi}}$, which follows directly from (7). Moreover, we find easier from (7) than from (2) that $\lim_{u \to -\infty} f_X(e^u) = f_X(0) = 0$ and that $f_X'(0) = 0$. This means that any lognormal starts at $t = 0$ from zero, increases up to the maximum at $t = e^{\mu - \sigma^2} > 0$ after which it decreases towards zero at $t \to \infty$. Thus, the lognormal is bell-shaped, but, in contrast to the Gaussian, the lognormal pdf is not symmetric around its maximum at $t = e^{\mu - \sigma^2}$ and can be seriously skewed.

The expression for the lognormal pdf in (2) can be rewritten [49] in a "power-law"-like form as

$$f_X(t) = \frac{e^{-\frac{\mu^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}} t^{-\alpha(t)} \tag{8}$$

where the exponent $\alpha(t)$ equals

$$\alpha(t) = 1 + \frac{\log t - 2\mu}{2\sigma^2} \tag{9}$$

which illustrates that a lognormal random variable behaves as a power-law random variable, provided the last fraction in (9) is negligibly small, say ε. The latter happens when $\left|\frac{\log t - 2\mu}{2\sigma^2}\right| < \varepsilon$. Thus, when $t \in [e^{2\mu - 2\sigma^2 \varepsilon}, e^{2\mu + 2\sigma^2 \varepsilon}]$ or in terms of the lower bound $\tau = e^{2\mu - 2\sigma^2 \varepsilon}$ and upper bound $\kappa = e^{2\mu + 2\sigma^2 \varepsilon}$ defined in Section 1, the pdf (2) of a lognormal random variable is almost indistinguishable from the pdf (1) of a power-law random variable with power exponent $\gamma \approx 1 + \varepsilon$. The t-interval $[e^{2\mu - 2\sigma^2 \varepsilon}, e^{2\mu + 2\sigma^2 \varepsilon}]$ exceeds the maximum $e^{\mu - \sigma^2}$ of the lognormal pdf and is clearly longer when σ is larger (as well as the tolerated accuracy ε increases). Malvergne *et al.* [49] mention that, if $\sigma = 3.4$, then the exponent $\alpha(t)$ in (9) varies less than 0.3 units over a range of three orders of magnitude.

Taking the logarithm on both sides of (2) results in

$$\ln(f_X(t)) = -\frac{1}{2\sigma^2}\ln(t)^2 + (\frac{\mu}{\sigma^2} - 1)\ln(t) - \ln(\sqrt{2\pi}\sigma) - \frac{\mu^2}{2\sigma^2} \qquad (10)$$

If σ is large, then the second term $(\frac{\mu}{\sigma^2} - 1)\ln(t)$ in (10) dominates, which leads to a straight line with in a log-log plot resembling a power-law with exponent $\gamma = 1$. On the other hand, if σ is small or $\mu = \sigma^2$, the first, quadratic term in (10) dominates.

B Distinguishing power-law from lognormals by likelihood testing

We apply the method of Clauset *et al.* [20] and Virkar and Clauset [51] to our data to distinguish between power-law and lognormal distributions. Clauset *et al.* [20] approached the problem of estimating the exponent γ in (1) by testing different distributions. In their data, lognormals and power-laws were not clearly distinguishable either: for some tested data-sets, a lognormal distribution actually achieved a higher p-values (goodness of fit) than power-laws, but log ratio tests suggested that other tested distributions are closer to power-laws. By using the technique in [20] to fit the distribution of T_{friend}, a power-law with exponent $\gamma = 1.53$ was found with a reasonable *p*-value of 0.23. The distribution of $T_{follower}$ is most likely not a power-law because the *p*-value of 0.0. Table 3 lists the parameters for all used datasets using Clauset and Virkar's fitting technique [51] with the according log-likelihood (Log-lh) and p values. A positive log-likelihood ratio indicates that the power-law is favored over the lognormal.

Distribution (binned in)	power-law Exponent (γ)	p	xmin	lognormal μ	σ	Log-lh p	LR	# of data-points
adding friends T_{friend} (seconds)	1.53	0.23	59	0.54	2.26	0.0015	418.67	7,156,722
adding follower $T_{follower}$ (seconds)	2.03	0.0	12	10.45	2.75	0	-1327.5	6,734,405
adding follower $T_{follower}$ (hours)	1.81	0.37	116	6.43	3.1	0	-297.66	41,184
sending emails (seconds)	2.33	0.0	86	9.53	3.24	0	-373.7441	68,346,901
sending emails (minutes)	2.03	0.21	1	6.9	3.32	0	-391.13	62,031,701
receiving emails (seconds)	2.01	0.0	1	10.31	2.73	0	-958.67	69,528,701
receiving emails (minutes)	2.21	0.05	182	6.23	3.54	0	-136.31	15,365,001
commenting on Reddit.com (seconds)	1.08	0.0	112	10.2	3.13	0	-0.0024	180,156,539
commenting on Reddit.com (hours)	1.91	0.01	64	0.88	2.41	0	-20.73	50,044

Table 3: Estimated parameters using the method of Clauset *et al.* [20]