Lognormal Infection Times of Online Information Spread

Christian Doerr*, Norbert Blenn, Piet Van Mieghem
Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, The Netherlands

Abstract
The infection times of individuals in online information spread such as the inter-arrival time of Twitter messages or the propagation time of news stories on a social media site can be explained through a convolution of lognormally distributed observation and reaction times of the individual participants. Experimental measurements support the lognormal shape of the individual contributing processes, and have resemblance to previously reported lognormal distributions of human behavior and contagious processes.

Introduction
The analysis of human social dynamics stemming from the emergent effects of individual human interactions has recently created a spur of research activity. Encompassing a wide area ranging from the propagation of opinions, epidemic spreading of information and innovation across groups of individuals [1], human dynamics investigates the temporal dependencies and interaction characteristics of the underlying contact and spreading processes, which are known to significantly drive overall diffusion and determine its contagion [2].

The analysis of the observed temporal distributions of online human activity data such as the inter-arrival and forwarding times of email [3], the propagation time of microblog posts [4], or telephony holding times [5], have commonly been approached and modeled [6] similar with the techniques established in topological network analysis, for example the approximation by power-law distributions. The applicability of fitting temporal behavioral data by a power-law has however been questioned [7,8] and bears a number of complications. First, the approximation through a power-law primarily concentrates on the fat tail providing a sub-optimal fit for the lower part of the observations, a power-law would imply that the highest frequency of data points is close to or equal to zero, yet our measurements show an up-going power-law would imply that the highest frequency of data points is close to or equal to zero, yet our measurements show an up-going

While a normal distribution describes well-behaved events, characterized by a mean and a standard deviation, such as the height of men and the temperature at noon in summer, the lognormal distribution exponentially blows up the relatively controlled deviations around the mean. Second, as a consequence of the Central Limit Theorem (CLT), the product of many independent identically distributed (i.i.d) random variables tends for large $n$ to a lognormal distribution. Slightly more general multiplicative processes (under certain conditions), such as the law of proportionate effect and geometric Brownian motion, give rise to lognormal behavior. Third, Marlow [18] beautifully demonstrates that, if a properly scaled sum of $n$ random variables tends to a normal distribution $N(0,1)$ for large $n$, the logarithm of the sum (also properly scaled) also tends to $N(0,1)$. In other words, both the scaled sum and the scaled logarithm of the sum converge to a same Gaussian $N(0,1)$, though at different rate. Yet differently stated, if the scaled sum of random variables tends to a Gaussian (by virtue of the CLT), then that same sum, though differently scaled, tends to a lognormal. Marlow’s theorem proves the observation in radio communication that a sum of lognormal random variables also

\[ f_Y(y) = \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{(\log y - \mu)^2}{2\sigma^2}} \]
tends to a lognormal. These limit laws illustrate why lognormals may appear relatively frequently.

In this paper, we study the inter-arrival time of retweets in Twitter and the spreading times on Digg. Both times reflect human interaction through communication technology. We show that this relatively new type of human interaction is lognormally distributed.

The Distribution of the Spreading Time $T$

In order to spread information, such as forwarding a message or news item on a social media platform, three consecutive processes take place as shown in figure 2: First, after an e-mail, tweet or message has been sent by a person, the information is processed and physically copied across the network of servers of the large social media sites and is delivered into the inboxes and queues of the receiving users where they compete for the user’s attention [19]. This first action requires $D_n$ time units, which we call the network propagation time. Second, users have to become aware of the content, for example by logging into the platform. The time between delivery and observation is denoted as the observation time $D_o$. Third, users decide to actively spread the information, for example through actions such as “retweets”, “likes”, “diggs”, thereby effectively forwarding the message to their connected friends and/or followers. The time between observation and passing a message is the reaction time $D_r$. The overall person-to-person forwarding time $T$ is the sum of these three time components: $T = D_n + D_o + D_r$.

In the following discussion, time measurements from the microblog service Twitter and the (former) social news aggregator Digg.com are used to explore the spread of information on online social media. For the case of Twitter we measure $T$ as the time to forward (“retweet”) messages by the followers of tweet originators, on Digg.com $T$ is denoted as the time between a user’s recommendation (called a “digg”) and the resulting diggs of a person’s followers. For both these services, users see a summary of their friends’ activities after visiting and logging into the website [20], and are presented with the opportunity to “retweet” or “digg” next to a particular piece of information.

Measurements of the network time $D_n$ indicated that $D_n$ is about two orders of magnitude smaller than the observation and reaction times. The network effect can therefore be neglected and the overall person-to-person forwarding time can be approximated by $T \approx D_o + D_r$. Hence, $T$ is, to a good approximation, determined by the habits and behaviors of the human participants.

In the following, we make two basic hypotheses:

1. We assume that the random variable $D_r \geq 0$ is independent of $D_o \geq 0$, i.e., the time to react to a message does not depend on the observation time, which allows the probability density function of $T = D_o + D_r$ to be expressed as a convolution of those of $D_o$ and $D_r$:

   \[
   f_{D_o + D_r}(t) = f_{D_o}(t) * f_{D_r}(t)
   \]  

   \[
   = \int_0^t f_{D_o}(x)f_{D_r}(t-x)dx
   \]

2. We make the hypothesis based on previous findings ([3,13,21], see discussion) that the observation and reaction times are lognormally distributed, $D_o \sim \logn(\mu_o, \sigma_o)$ and $D_r \sim \logn(\mu_r, \sigma_r)$, so that with (1),
Figure 3. Distribution of spreading times on Twitter.
doi:10.1371/journal.pone.0064349.g003
upvoted a news item, and this recommendation will be displayed on the follower’s user interface on his next login on the website. The reaction time equals the interval between the notification and the receiver’s upvote (if applicable).

Since the public record of diggs and comments does not directly list the instances at which a user has consulted his inbox to see the incoming notifications during a particular visit on the website, this monitoring procedure will only create a general bound of the observation and reaction times (see figure 7). As the notification

Figure 4. Distribution of spreading times on Digg.
doi:10.1371/journal.pone.0064349.g004

Figure 5. Error of the lognormal fit for Twitter of Figure 3.
doi:10.1371/journal.pone.0064349.g005
page is only visible after a login (and the login event must have been completed by the time we observed the first digg), the time between the original digg and the approximated login time of the follower will be lower bound of the observation time, provided that a particular person could have checked the incoming recommendations at any time point after the login. Correspondingly, the time between the login event and the person’s own digg will therefore be an upper bound of the reaction time, as the notification triggering the digg could have been read anytime between the login and the actual action. The same approach is followed in case of Twitter, where the timestamps of tweets will be used to establish an estimate when a user is present on the social network, leading upper and lower bounds of reaction and observation time respectively.

The analysis of the experimental data shows that both lower and upper bound of observation and reaction time are following a lognormal distribution, thereby confirming our initial hypotheses for the spreading time $T$. Table 1 list the ML-fit based on (4) against the experimentally obtained parameters of the underlying distributions. The two distributions from the maximum-likelihood fits also line up on the correct side and neighborhood of the empirical lower and upper bounds, thereby making it possible to identify $m_1, \sigma_1$ as the reaction time (colored red in the insets of figures 1, 3 and 4), and $m_2, \sigma_2$ as the observation time (colored orange in the insets). The measured and fitted observation time parameters $m_0$, $\sigma_0$ match well those predicted from our earlier maximum-likelihood fit. The mean of the measured reaction time also falls within the expected interval, between 0 and the computed upper bound, although the bound is less tight.

### Materials and Methods

The analysis of spreading times was based on two datasets, a collection of initial original posts and their corresponding retweets on the microblog Twitter, and five year trace of stories and votes submitted to the social media aggregator Digg.com. Both datasets were collected from the providers’ webpages and API interfaces and contain only publicly available information. The former dataset of 20.5 million tweets was constructed in a two step collection of initial original posts and their corresponding retweets. The latter dataset is comprised of the content submissions to the social media aggregator Digg.com within the seven year period of 2004–2010 and the recommending votes of its users. Using the crawling technique described in [20], we obtained a complete record of the 310 million votes from 2.2 million users’ activity on the 12 million submitted stories.

### Discussion

The analysis and the experimental data point to lognormal distributions as an explanation of the temporal distributions observed in epidemic spreading online. This finding is not per se surprising as lognormal distributions have for a long time been observed in and connected to human behavior, which is also driving the spread of information and innovation online.

The analogue to the reaction time $D_r$ is, for example, in classical epidemiology referred to as the incubation period. This incubation time in epidemic processes between the infection of an individual and the first symptoms or active transmission of the disease is frequently found to adhere to a lognormal distribution, for example in the case chicken pox, hepatitis, or salmonellosis [12, 23]. In an overview of incubation periods of 86 diseases, Nishiura concludes that 70.9% can be accepted as lognormal at a 5% level of significance [21].

Similar lognormal patterns have been discovered across domains for the equivalent of the observation time $D_o$. Barcelo [24] reports the channel holding time on cellular networks, the duration of calls and thus the time of the transmission process to be lognormally distributed.

This general notion of the lognormally distributed duration of human activities and associated spreading tasks are also repeatedly found across other domains, for example in the duration of strikes [14] or the time to complete tasks on a test [13]. This same pattern seems to extend into the online domain for spreading times, for

| Table 1. Comparison of observation and reaction times of ML-fit and derived bounds from Digg and Twitter dataset. |
|---------------------------------------------------------------|
| **Measurement**                      | **Digg** | **Twitter** |
|-------------------------------------|----------|-------------|
| Lognormal $D_o$ (ML-Fit)            | 10.2     | 5.6         |
| Lower Bound of Observation Time     | 9.48     | 5.5         |
| Lognormal $D_r$ (ML-Fit)            | 2.1      | 3.7         |
| Upper Bound of Reaction Time        | 6.26     | 6.92        |

[10.1371/journal.pone.0064349.t001]
The lognormal distribution regularly appears in the broader context of universal human activities and behavior. When tracing the mobility of cell phone handsets, Barcelo and Jordan [5] find that the time a person stays at a certain location (and is associated with a particular cell phone call) also follows a lognormal distribution, which at near 100% penetration in many countries can be seen as a good proxy to measure human mobility. Radicchi, Fortunato and Castellano [25] for example report that the citation counts of academic articles universally follows a lognormal distribution when normalized by the relative citation habits in a scientific field. A similar phenomenon can also be observed in the law of proportionate effect [15], demonstrating that the publicly visible total number of recommendations changes the voting behavior of individuals in Digg and leads at a population-level to results best characterized by the lognormal distribution [16].

Similarly, a variety of such lognormally distributed properties of populations have been discovered, such as in the distribution of incomes or marriage ages [12]. Shockley [26] explains the lognormal distribution of the duration of strikes. A normal limit theorem for power sums of normal random variables. The Bell System Technical Journal 46: 2081–2089.

The lognormal distribution is a probability distribution that is often used to model data that are always positive and have a large range of values. It is often used to model natural phenomena such as the distribution of income or the distribution of wealth.

References

1. Kitsak M, Gallos LK, Havlin S, Liljeros F,Muchnak L, et al. (2010) Identification of incipient spreaders in complex networks. Nature Physics 6: 888–893.
2. Vespignani A (2009) Predicting the behavior of techno-social systems. Science 325.
3. Iribarren JL, Moro E (2009) Impact of human activity patterns on the dynamics of information diffusion. Physical Review Letters 103.
4. Yu LL, Asur S, Huberman BA (2012) Artificial ination: The real story of trends and trend-setters in sina weibo. In: SocialCom.
5. Barcelo F, Jordan J (2000) Channel holding time distribution in public telephony systems (pamr and pcs). IEEE Transactions on Vehicular Technology.
6. Vajna S, Toth B, Kertész J (2012). Modelling power-law distributed interevent times. arXiv:1211.1173.
7. Mahgreen RD, Stouffer DB, Motter AE, Amaral LAN (2006) A poissonian explanation for heavy tails in e-mail communication. PNAS 105.
8. Stouffer DB, Mahgreen RD, Amaral LAN (2005). Comment on the origin of bursts and heavy tails in human dynamics. arXiv:physics/0510216.
9. Bollobas B, Riordan O, Spencer J, Tusnady G (2001) The degree sequence of a scale-free random graph process. Random Structures & Algorithms 18: 279–290.
10. Barabasi AL, Albert R (1999) Emergence of scaling in random networks. Science 286.
11. Crow EL, Shimizu K (1968) Lognormal Distribution, Theory and Application.
12. Limpert E, Stahel WA, Abbt M (2001) Log-normal distributions across the sciences: Keys and clues. Bio Science 51.
13. van der Linden VJ (2006) A lognormal model for response times on test items. Journal of Educational and Behavioral Statistics 31.
14. Lawrence RJ (1984) The lognormal distribution of the duration of strikes. Journal of the Royal Statistical Society Series A 147.
15. Gibar R (1931, French) Les Inégalités économiques [The Economic Inequalities]. Librairie du Recueil Sirey.
16. Van Mieghem P, Blenn N, Doerr C (2011) Lognormal distribution in the digg online social network. EPJ B 83: 251–261.
17. Van Mieghem P (2006) Performance Analysis of Communications Systems and Networks. Cambridge University Press.
18. Marlow NA (1967) A normal limit theorem for power sums of normal random variables. The Bell System Technical Journal 46: 2081–2089.
19. Gonçalves B, Perra N, Vespignani A (2011) Modeling users’ activity on twitter networks: Validation of dunbar’s number. PLoS ONE 6: e22656.
20. Doerr C, Blenn N, Tang S, Van Mieghem P (2012) Are friends overrated? A study for the social news aggregator digg.com. Computer Communications 35.
21. Nishiura H (2007) Early efforts in modeling the incubation period of infectious diseases with an acute course of illness. Emerging Themes in Epidemiology.
22. Newman MEJ (2005) Power laws, pareto distributions and zipf’s law. Contemporary Physics.
23. Sartwell PE (1950) The distribution of incubation periods of infectious diseases with an acute course of illness. American Journal of Hygiene 51: 310–318.
24. Barcelo F (1998) Channel holding time distribution in public cellular telephony. Electronic Letters 34.
25. Radicchi F, Fortunato S, Castellano C (2008) Universality of citation distributions: Toward an objective measure of scientific impact. PNAS 105.
26. Shockley W (1957) On the statisticsof individual variations of productivity in research laboratories. Proceedings of the ERE.
27. Vazquez A, Oliveira JG, Dezo Z, Goh KI, Kondor I, et al. (2006) Modeling bursts and heavy tails in human dynamics. Physical Review E 73.

Conclusion

Based on a huge number of measurements in Twitter and Digg, we found that human related activities such as the observation time $D_o$ and reaction time $D_r$ are very close to a lognormal random variable. While favoring power-law behavior, Vazquez et al. [27] argue that a proper identification of a lognormal (and its distinction from the ubiquitously present power-law) distribution requires the availability of extensive datasets. With an analysis of two datasets of 310 and 20.5 million interactions, we believe that the presented verification here amply confirms the lognormal behavior. Although there is theoretical support for the occurrence of lognormal random variables, the precise underlying stochastic processes, that lead to a lognormal outcome are today still missing.

Acknowledgments

This research has been supported by the EU Network of Excellence in Internet Science EINS (project no. 288021) and the Trans-sector Research Academy for Complex Networks and Services (TRANS).

Author Contributions

Conceived and designed the experiments: CD NB. Performed the experiments: CD NB. Analyzed the data: CD NB PVM. Wrote the paper: CD NB PVM.

Figure 7. A lower bound estimate of the observation time and an upper bound of the user’s reaction times can be determined when augmenting the voting behavior with the friendship network. doi:10.1371/journal.pone.0064349.g007