EPJ B
Condensed Matter and Complex Systems
EPJ.org
your physics journal

Eur. Phys. J. B **83**, 251–261 (2011) DOI: 10.1140/epjb/e2011-20124-0

Lognormal distribution in the digg online social network

P. Van Mieghem, N. Blenn and C. Doerr

EDP SCIENCES
Società Italiana di Fisica
Springer

Eur. Phys. J. B **83**, 251–261 (2011)
DOI: 10.1140/epjb/e2011-20124-0

THE EUROPEAN PHYSICAL JOURNAL B

Regular Article

Lognormal distribution in the digg online social network

P. Van Mieghem[a], N. Blenn, and C. Doerr

Delft University of Technology, Faculty of EEMCS, 2628 CD Delft, Netherlands

Received 16 February 2011 / Received in final form 2 June 2011
Published online 15 September 2011 – © EDP Sciences, Società Italiana di Fisica, Springer-Verlag 2011

Abstract. We analyse the number of votes, called the digg value, which measures the impact or popularity of submitted information in the Online Social Network Digg. Experiments over five years indicate that the digg value of a story on the first frontpage follows closely a lognormal distribution. While the law of proportionate effect explains lognormal behavior, the proportionality factor a in that law is assumed to have a constant mean, whereas experiments show that a decreases linearly with time. Our hypothesis, *the probability that a user diggs (votes) on a story given that he observes a certain digg value m equals $a \times m$*, can explain observations, provided that the population of users that can digg on that story is close to a Gaussian.

1 Introduction

In recent years, online social networks (OSNs) have experienced an explosive growth. With hundreds of such services available and a subscribed user base of several hundred million people, they have significantly altered the way how people spend their time [1] and how they search for content [2].

One specific kind of OSNs addressed in this paper are social bookmarking services and, in particular, social news aggregators such as digg.com, delicious.com or reddit.com. In this type of OSN, users are sharing and commenting on information (such as bookmarks, opinions, news, etc.), further called "stories". The community will vote (an activity referred to as *"to digg"*) on the submitted stories, where the sum of all votes on a story, which is called the "digg value" of a story, is publicly displayed as some ranking information and therefore reflects the impact of a story. The life-time of a story is also studied and used in [3] to "predict the popularity of online content". Collaborative tagging in social media, as in digg.com, has been overviewed in [4] from a statistical physics point of view.

Empirical data (see e.g. [5,6]) illustrate that the digg value of an arbitrary story follows a lognormal distribution as shown in Figure 1. For over hundred years, lognormal distributions have been observed in many different areas [7–9], from economy to biology and now in OSNs. Recent work [10] further demonstrated that a lognormal distribution may further universally characterize datasets which have previously been thought of as typical instances of a power-law. The fascinating underlying process that asymptotically generates a lognormal distribution is the law of proportionate effect, which is briefly reviewed in Appendix A. Earlier, Wu and Huberman [11] have argued that the digg value is, indeed, generated by the law of proportionate effect when ageing effects are taken into account. However, Section 6 questions their arguments.

In this article, we propose another process in Section 3 that leads to a lognormal distribution: our counting process avoids asymptotic limits, but in return requires that the user population is approximately a Gaussian. The new insight is the probabilistic relation (12) in Section 3.2 of whether a user will digg on a story, given that he observes the digg value. The remainder of the paper tries to relate both governing processes with experiments. The mechanics of Digg and the data extraction method from the Digg OSN, observed over a period of five years, are described in [6,12]. Section 7 concludes our exploration: the law of proportionate effect alone seems insufficient to explain the experimental evidence, while our counting process is able to explain most findings, provided the user population is normally distributed. The key quantity is the proportionality factor a. Experiments exhibit a roughly linearly decreasing a with time. The average of the proportionality factor a in the law of proportionate effect (Appendix A) is, however, constant. Our finite counting process and the hypothesis (12) agree with experiments: a decreases linearly in the number of users that could digg on a story and approximately linearly with time.

Our results may have broader applicability, as [13] for example also find a log-normal behavior in the popularity of movies at various levels, and hypothesize a self-reinforcing process from explicit social recommendations as a potential cause.

2 The digg value of a story

Let D_s denote the digg value, the number of diggs, on story s. If we assume that a user can only digg once on a story, then

$$D_s = \sum_{j=1}^{\mathcal{N}_s} 1_{\{\text{user } j \text{ diggs on } s\}} \tag{1}$$

[a] e-mail: p.f.a.vanmieghem@tudelft.nl

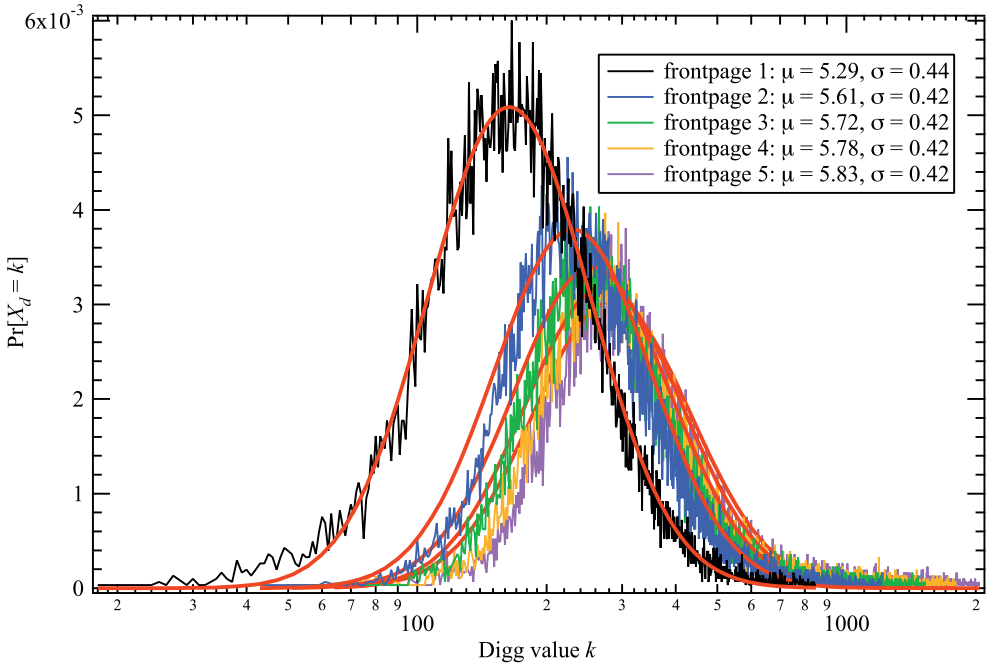

Fig. 1. (Color online) The probability density function (pdf) of the digg value X_d of stories on the first five frontpages in Digg. Each pdf is fitted by a lognormal (2) and the corresponding parameters μ and σ are shown in the legend.

where $\mathcal{N}_s$ is the total number of users that have the opportunity to digg on story s. The indicator function 1_x equals one if the event x is true, else it is zero.

2.1 Experiments

Empirical data illustrated in Figure 1 show that only the digg value of stories on the first frontpage is well fitted by a lognormal density function [14], p. 57

$$f_{\text{lognormal}}(u) = \frac{1}{u\sigma\sqrt{2\pi}} \exp\left[-\frac{(\log u - \mu)^2}{2\sigma^2}\right]. \quad (2)$$

The probability density function (pdf) of the other frontpages decays in the left- and right tail faster than a lognormal distribution. The decrease faster than a lognormal is likely due to the ageing of the story (explained in Sect. 2.2), the decreasing number of users that visit subsequential frontpages and the convolution effect (that forces any distribution towards a Gaussian distribution). The latter is a consequence of the definition $D_s = D_{s0} + D_{s1} + D_{s2} + \ldots$, where D_{sk} is the sum of the diggs on story s while on frontpage $k = 1, 2, \ldots$ and D_{s0} is the digg value of story s just before it is promoted from the upcoming section to the first frontpage. Although D_{sk} and D_{sm} are not independent, we infer that they are only weakly dependent.

2.2 Users digg dependently

It is conceivable that the probability that the j-th user, further called user j, diggs on story s is dependent on the digg value that he observes, when his eye catches the story s. The digg value that user j sees when first encountering the story s is, for $j > 1$, the sum of all digg values of users that dugg on the story before him/her

$$X_s(j) = \sum_{k=1}^{j-1} 1_{\{\text{user } k \text{ diggs on } s\}}. \quad (3)$$

The eventual digg value of story s is $D_s = X_s(\mathcal{N}_s + 1)$, while $X_s(1) = 0$ and $E[X_s(2)] = p_1$, the probability that the first user diggs on story s. We may argue that the digging probability of a user is influenced by the story's digg value: if the number of diggs is low, a user has an a priori feeling that the story is not so attractive and his motivation to digg is lowered, while the opposite occurs when a high number of diggs is observed.

Another phenomenon is ageing: after some time, the novelty of the story may diminish, especially if the story contains news or temporary information. In that case, the overall motivation to digg on the story, independent of the digg value, decreases. Wu and Huberman [11] explain that the story's growing attraction, measured via its increasing digg value, is counterbalanced by the ageing of the story. As discussed in Section 6, the analysis and assumptions of ageing effects in [11] are debatable. Moreover, the user population that visits the k-th frontpage for $k > 4$ is decreasing so quickly that data is scarce. Therefore, in the sequel, ageing is not considered, nor time-dependent effects that affect the digg value. We also ignore primacy and recency effects [15] that would predict that the top most and bottom most elements would attract larger attention, simply because of their position in a long list. For example, the Digg feature, "Top Stories in All Topics",

displays 10 stories in an abbreviated format on each frontpage, so that the list can be seen in one scan by any visitor (15 items on a VGA screen require one scrolling action). Due to their special placement alone (irrespective of their content), Top Stories are likely to receive a higher than "normal" number of diggs. In the sequel, we do not distinguish between stories.

Our main interest is to figure out whether the motivation to digg on a story is linearly dependent on that story's diggvalue. If a linear dependence holds, a lognormal distribution is the natural consequence by the law of proportionate effect (Appendix A). In order to keep external factors as constant as possible and guided by the experiments in Figure 1, we mainly concentrate on the number of digg values of a story as long as it is on the first frontpage.

3 A general description for the average digg value

We take advantage of the fact that the digg value, both $X_s(j)$ as D_s, is a sum of indicators: Bernoulli random variables that are either zero or one. Since $E\left[1_{\{\text{user } k \text{ diggs on } s\}}\right] = \Pr[\text{user } k \text{ diggs on } s]$ and since the expectation operator is linear (irrespective of dependencies between random variables in the sum), taking the expectation of (3) yields

$$E[X_s(j)] = \sum_{k=1}^{j-1} \Pr[\text{user } k \text{ diggs on } s].$$

By the law of total probability [14], we have that

$$\Pr[\text{user } k \text{ diggs on } s] = \sum_{m=0}^{k-1} \Pr[\text{user } k \text{ diggs on } s |\, X_s(k) = m] \Pr[X_s(k) = m].$$

When writing the conditional probability as

$$\Pr[\text{user } k \text{ diggs on } s |\, X_s(k) = m] = g_k(m) \qquad (4)$$

where $g_k(x)$ is a non-negative function that maps x to the interval $[0, 1]$, we obtain

$$\Pr[\text{user } k \text{ diggs on } s] = \sum_{m=0}^{k-1} g_k(m) \Pr[X_s(k) = m] = E[g_k(X_s(k))] \qquad (5)$$

where we have used the definition of the expectation of a function of a random variable [14], p. 17. Thus, we arrive at the general relation for the average number of diggs that the jth user sees

$$E[X_s(j)] = \sum_{k=1}^{j-1} E[g_k(X_s(k))] \qquad (6)$$

from which we obtain the difference equation

$$E[X_s(j)] - E[X_s(j-1)] = E[g_{j-1}(X_s(j-1))]. \qquad (7)$$

The average total number $D_s = X_s(\mathcal{N}_s + 1)$ of diggs on story s by a population of $\mathcal{N}_s$ users is

$$E[D_s] = \sum_{k=1}^{\mathcal{N}_s} E[g_k(X_s(k))].$$

We illustrate in Appendix B that, for positively correlated users as in Digg, the variance $\mathrm{Var}[X_s(j)]$ can be small. In fact, the stronger the correlation in the digging behavior between users, the smaller the variance. A small variance $\mathrm{Var}[X_s(j)]$ implies that the mean $E[X_s(j)]$ is a good approximation for the random variable $X_s(j)$.

So far, we have implicitly assumed that $\mathcal{N}_s$ is a constant. However, $\mathcal{N}_s$ is, in fact, also a random variable, denoting the number of users that has discovered story s within a certain time interval. If $\mathcal{N}_s$ is a random variable, then the above computation is valid for the conditional expectation $Y_s = E[D_s|\mathcal{N}_s]$, which is the random variable equal to the average number of diggs on a story s given that the total number of users equals $\mathcal{N}_s$.

We believe that the r.v. Y_s is approximately measured. Each story s has a number D_s of diggs that are recorded, while the total population of potential diggers is $\mathcal{N}_s$. Hence, we have a sequence of stories with their corresponding diggs and population $\{(D_s, \mathcal{N}_s)\}_{s\geq 1}$. Clearly, it follows from the definition (1) that $D_s \leq \mathcal{N}_s$. If the information about the population is omitted, intuitively one feels that D_1 cannot be compared to D_2, because it is obvious that if $n_2 = \frac{n_1}{2}$, then the number of diggs D_2 cannot be higher than $\frac{D_1}{2}$. Yet, all digg values of the different stories, the sequence $\{D_s\}_{s\geq 1}$, are placed in 1 histogram. In fact, by doing so, we compare different processes, while a histogram should only record the outcomes of a same stochastic process and each outcome should be independent of all others. Therefore, we believe that the histogram approximates the conditional random variable $D_s|\mathcal{N}_s$ by its best guess, the estimated value $E[D_s|\mathcal{N}_s] = Y_s$.

The remainder consists of choosing the function $g_k(x)$, defined in (4).

3.1 A simple case: proportionality

A simple mathematical choice is the linear function $g_k(x) = a_k x + b_k$ and the general difference (7) becomes

$$E[X_s(j)] = (1 + a_{j-1}) E[X_s(j-1)] + b_{j-1}.$$

After m iterations of this difference equation, we obtain

$$E[X_s(j)] = E[X_s(j-m)] \prod_{k=1}^{m} (1 + a_{j-k}) + \sum_{l=1}^{m} b_{j-l} \prod_{k=1}^{l-1} (1 + a_{j-k}) \qquad (8)$$

and for $j = \mathcal{N}_s + 1$ since $D_s = X_s(\mathcal{N}_s + 1)$,

$$E[D_s] = E[X_s(\mathcal{N}_s + 1 - m)] \prod_{k=\mathcal{N}_s+1-m}^{\mathcal{N}_s} (1 + a_k) + \sum_{l=\mathcal{N}_s+1-m}^{\mathcal{N}_s} b_l \prod_{k=l+1}^{\mathcal{N}_s} (1 + a_k).$$

With the initial condition $E[X_s(2)] = p_1$, we have

$$E[D_s] = p_1 \prod_{k=2}^{\mathcal{N}_s} (1 + a_k) + \sum_{l=2}^{\mathcal{N}_s} b_l \prod_{k=l+1}^{\mathcal{N}_s} (1 + a_k). \quad (9)$$

When choosing all $a_k = a$ and all $b_k = b$, then expression (8) simplifies to

$$E[X_s(j)] = E[X_s(j - m)](1 + a)^m + b \sum_{l=1}^{m} (1 + a)^{l-1} = \left(E[X_s(j - m)] + \frac{b}{a}\right)(1 + a)^m - \frac{b}{a} \quad (10)$$

such that, for $j = \mathcal{N}_s + 1$,

$$E[D_s] = \left(E[X_s(\mathcal{N}_s + 1 - m)] + \frac{b}{a}\right)(1 + a)^m - \frac{b}{a}. \quad (11)$$

Hence, the average digg value $E[D_s] = E[D_{s1}]$ of a story just before it disappears from the frontpage can be expressed as a function of the average digg value $E[X_s(\mathcal{N}_s + 1 - m)] = E[D_{s0}]$ just after the story appears on the frontpage[1]. During the time on the frontpage, precisely m users have had the opportunity to digg on that story.

Suppose that the number of users m in (11) is a Gaussian $N(\widetilde{\mu}, \widetilde{\sigma}^2)$, then the random variable $Y_{s1} = E[D_{s1}|m]$ is a lognormal, provided $E[X_s(\mathcal{N}_s + 1 - m)] = E[D_{s0}]$ is a known constant. Indeed,

$$\Pr[Y_{s1} \le x] = \Pr\left[\left(E[D_{s0}] + \frac{b}{a}\right)(1 + a)^m - \frac{b}{a} \le x\right] = \Pr\left[m \le \frac{\log \frac{x+\frac{b}{a}}{E[D_{s0}]+\frac{b}{a}}}{\log(1 + a)}\right] = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\frac{\log \frac{x+\frac{b}{a}}{E[D_{s0}]+\frac{b}{a}}}{\log(1+a)}} \exp\left[-\frac{(t - \widetilde{\mu})^2}{2\widetilde{\sigma}^2}\right] dt.$$

such that

$$f_{Y_{s1}}(x) = \frac{d \Pr[Y_{s1} \le x]}{dx} = \frac{1}{\left(x + \frac{b}{a}\right)\widetilde{\sigma}\log(1 + a)\sqrt{2\pi}} \times \exp\left[-\frac{(\log\left(x+\frac{b}{a}\right) - \log\left(E[D_{s0}]+\frac{b}{a}\right) - \widetilde{\mu}\log(1+a))^2}{2\widetilde{\sigma}^2 \log^2(1 + a)}\right]$$

which is recognized from (2) as a lognormal distribution in $u = x + \frac{b}{a}$ with parameters $\mu = \log\left(E[D_{s0}] + \frac{b}{a}\right) + \widetilde{\mu}\log(1 + a)$ and $\sigma = \widetilde{\sigma}\log(1 + a)$. Hence, assuming a Gaussian population, the digg value D_{sk} of the story on frontpage k can be described by a lognormal distribution, possibly with different parameters a and b when ageing is taken into account.

Clearly, the expressions simplify considerably if $b = 0$. In the case $b = 0$, our derivation is in line with the law of proportionate effect as shown in Section 4. However, the parameters of the lognormal distribution are different.

3.2 Interpretation

The choice $g_j(x) = a_j x + b_j$ implies, as follows from (4), that

$$\Pr[\text{user } j \text{ diggs on } s | X_s(j) = m] = a_j m + b_j.$$

This general linear form dependent on user j can be useful to specify the digging behavior of friends and non-friends of the originator of the story s. In case that user j is a friend of the originator of the story s, he/she may be insensitive to the digg value $X_s(j) = m$ that he/she observes when first encountering the story, because his/her friendship relation with the originator outweighs the judgements of other diggers. Hence, $a_j = 0$ and $b_j = b \le 1$, a constant value that expresses the faith or depth of the friendship relation of friends in the originator. On the other hand, non-friends have almost no faith in the originator, but in their peers, such that $a_j = a \le 1$ and $b_j = 0$. Consequently, the general linear relation (9) becomes

$$E[D_s] = p_1(1 + a)^{(\mathcal{N}_{\text{non-friends}}-2)} + b(\mathcal{N}_{\text{friends}} - 2)$$

where $\mathcal{N}_{\text{friends}}$ is the number of friends of the originator of the story s and $\mathcal{N}_s = \mathcal{N}_{\text{non-friends}} + \mathcal{N}_{\text{friends}}$. Assuming that both populations of friends and non-friends are Gaussian-like distributed (with possibly different mean and variance), this expression shows that $E[D_s | \mathcal{N}_{\text{friends}}, \mathcal{N}_{\text{non-friends}}]$ is the sum of a Gaussian and a lognormal, which is again heavy-tailed. Since $\mathcal{N}_{\text{friends}}$ is usually smaller than $\mathcal{N}_{\text{non-friends}}$, a lognormal distribution is expected to dominate [6].

When returning to the simplest linear case for $g_k(x) = a_j x + b_j$, where $a_j = a$ and $b_j = 0$, then

$$\Pr[\text{user } j \text{ diggs on } s | X_s(j) = m] = am \quad (12)$$

[1] Similarly, we can express the average number on the k-th frontpage by considering $E[D_s] = E[D_{sk}]$, $E[X_s(\mathcal{N}_s + 1 - m)] = E[D_{s,k-1}]$ and m is the number of users that had the opportunity to digg on the story during its stay on the k-th frontpage.

implying that $a > 0$ is assumed to be constant for each user j. Thus, $a \leq \frac{1}{\mathcal{N}_s}$ in order that all conditional probabilities, for each $j \leq \mathcal{N}_s$, are smaller than or equal to 1. Thus, (12) illustrates that a decreases with $\mathcal{N}_s$, which is in line with Figure 4. Furthermore, relation (5) reduces to

$$\Pr[\text{user } k \text{ diggs on } s] = aE[X_s(k)].$$

If the user is the k-th user, counted since the story is on the j-th frontpage, then, with (10), we have that

$$\Pr[\text{user } k \text{ diggs on } s] = aE[D_{s,j-1}](1+a)^k$$

which shows an exponentially increasing digging probability in k.

4 Application of the law of proportionate effect to the digg value

The law of proportionate effect, explained in Appendix A, is more general than the analysis in Section 2.2: it applies to the random variable $X_s(n)$ instead of the mean $E[X_s(n)]$. On the other hand, it is an asymptotic result $(n \to \infty)$, which implicitly assumes a rapid convergence towards a Gaussian in order to observe for finite n already the lognormal distribution. In our case, where the digg value $X_j = X_s(j+1)$ is a sum of indicators as defined in (3), we have that

$$X_s(j+1) - X_s(j) = 1_{\{\text{user } j \text{ diggs on } s\}}.$$

According to the law of proportionate effect $X_j = (1+\alpha_j)X_{j-1}$, we deduce that

$$1_{\{\text{user } j \text{ diggs on } s\}} = \alpha_j X_s(j). \tag{13}$$

Since α_j and $X_s(j)$ are independent as assumed in the law of proportionate effect, taking the expectation leads to

$$\Pr[\text{user } j \text{ diggs on } s] = E[\alpha_j]\,E[X_s(j)]$$

and we conclude from (5) with $g_k(x) = a_j x$ that $E[\alpha_j] = a_j$. The law of proportionate effect assumes that all random variables α_j are i.i.d. with mean $E[\alpha]$, such that $E[\alpha_j] = E[\alpha]$ and $a_j = a$ for each j. Hence, the conditional probability (12) is a manifestation of the law of proportionate effect and provides another way to verify proportional behavior.

We will now compute the parameter $\mu = E[\log(1+\alpha)]$ in the lognormal limit law for $X_n = X_s(n+1)$ for large n (see Appendix A). Taking the m-th power of (13) yields

$$1_{\{\text{user } j \text{ diggs on } s\}} = \alpha_j^m X_s^m(j)$$

from which it follows that

$$\begin{aligned}\Pr[\text{user } j \text{ diggs on } s] &= E[\alpha_j^m]\,E[X_s^m(j)] \\ &= E[\alpha^m]\,E[X_s^m(j)].\end{aligned}$$

Since $E[X_s(j)] > 1$ for not too small j, we may conclude that $E[\alpha^m] < 1$ for all $m \geq 1$. Hence, the series

$$\begin{aligned}E[\log(1+\alpha)] &= E\left[\sum_{m=1}^{\infty}(-1)^{m-1}\frac{\alpha^m}{m}\right] \\ &= \sum_{m=1}^{\infty}(-1)^{m-1}\frac{E[\alpha^m]}{m} \\ &= \Pr[\text{user } j \text{ diggs on } s]\sum_{m=1}^{\infty}\frac{(-1)^{m-1}}{mE[X_s^m(j)]}\end{aligned}$$

converges and the parameter $\mu = E[\log(1+\alpha)]$ equals

$$\mu = \Pr[\text{user } j \text{ diggs on } s]\sum_{m=1}^{\infty}\frac{(-1)^{m-1}}{mE[X_s^m(j)]}.$$

The latter alternating series with decreasing terms is bounded by

$$\frac{1}{E[X_s(j)]} - \frac{1}{2E[X_s^2(j)]} < \sum_{m=1}^{\infty}\frac{(-1)^{m-1}}{mE[X_s^m(j)]} < \frac{1}{E[X_s(j)]}.$$

Invoking the bounds yields

$$\begin{aligned}\frac{\Pr[\text{user } j \text{ diggs on } s]}{E[X_s(j)]}\left(1 - \frac{E[X_s(j)]}{2E[X_s^2(j)]}\right) &< \mu \\ &< \frac{\Pr[\text{user } j \text{ diggs on } s]}{E[X_s(j)]}\end{aligned}$$

and

$$\mu E[X_s(j)] < \Pr[\text{user } j \text{ diggs on } s] < \frac{E[X_s(j)]\,\mu}{\left(1 - \frac{E[X_s(j)]}{2E[X_s^2(j)]}\right)}.$$

Hence, we find that $\mu E[X_s(j)] < 1$. However, experimental results (Fig. 1) indicate that both $\mu \simeq 5.2$ and $E[X_s(j)]$ are larger than 1, contradicting $\mu E[X_s(j)] < 1$. The analysis indicates that the law of proportionate effect only sets in when a certain value of $X_s(j)$, or equivalently j, is reached.

This conclusion is supported by the analysis in Section 3.1 (and Fig. 3 below). Using (11), assuming that $E[X_s(\mathcal{N}_s + 1 - m)] = D_{s0}$, the digg value of a story s just before it is promoted to the frontpage, is a known constant and m is normally distributed as $N(\widetilde{\mu}, \widetilde{\sigma}^2)$, we obtain a lognormal with mean

$$\mu = \log\left(D_{s0} + \frac{b}{a}\right) + \widetilde{\mu}\log(1+a)$$

and $a \leq \frac{1}{m}$. Confining to the case where $b = 0$ and using the average $\widetilde{\mu}$ as a good estimate for m such that $a = \frac{1}{\widetilde{\mu}}$, then

$$\begin{aligned}\widetilde{\mu}\log(1+a) &\approx \widetilde{\mu}\log\left(1 + \frac{1}{\widetilde{\mu}}\right) \\ &= \widetilde{\mu}\left(\frac{1}{\widetilde{\mu}} - \frac{1}{2\widetilde{\mu}^2} + O\left(\frac{1}{\widetilde{\mu}^3}\right)\right) \\ &= 1 - \frac{1}{2\widetilde{\mu}} + O\left(\frac{1}{\widetilde{\mu}^2}\right) < 1\end{aligned}$$

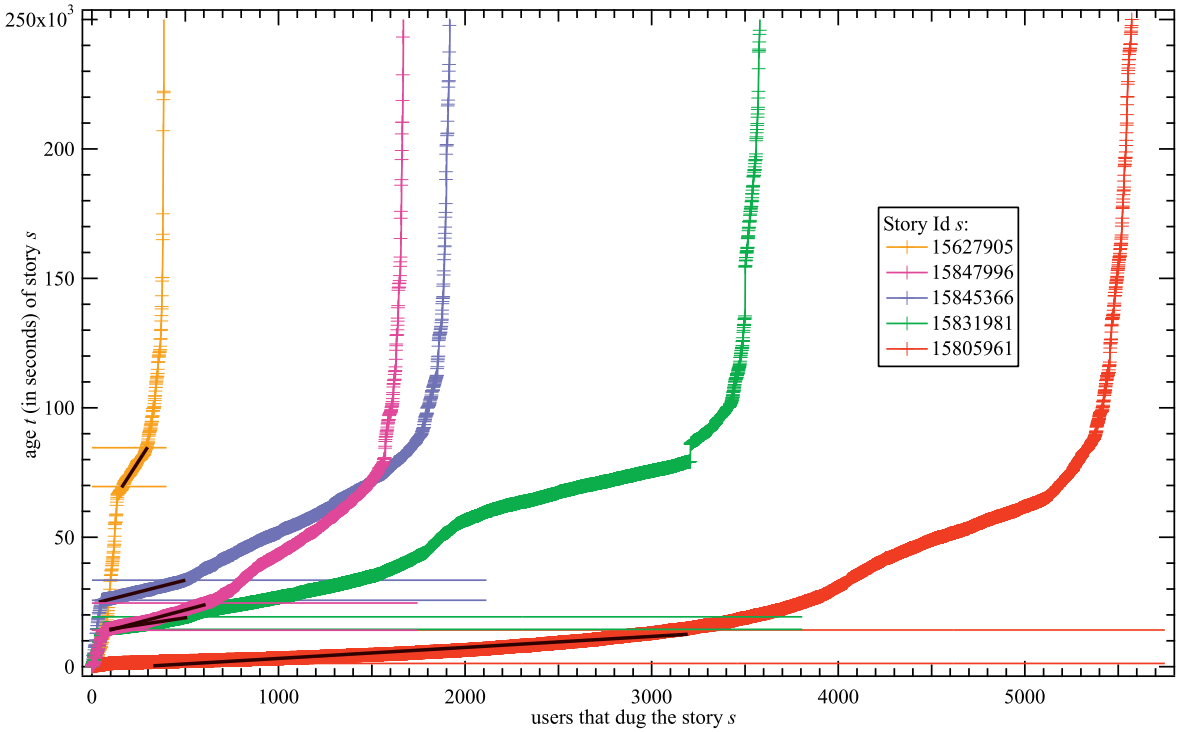

Fig. 2. (Color online) The time t of a story s on which a user has dug on that story. Five typical stories are shown. The horizontal lines indicate the duration when the story was on the frontpage and the time t seems to be approximately linear with the the number of digging users while the story is on the frontpage. A story stays, on average, 2.4 h on the first frontpage, on which most Digg users are active.

and

$$\mu - \log\left(D_{s0}\right) \lessapprox 1.$$

Since $\mu \simeq 5.2$ in Figure 1, we deduce that $D_{s0} \gtrapprox e^{4.2}$, which agrees with experiments [6].

5 The incremental increase of the digg value with time

If the law of proportionate effect is correct, then it follows from (A.2) that

$$X_n - X_m = X_m \left(\prod_{j=m+1}^{n} (1+\alpha_j) - 1 \right).$$

For not too small m, (13) shows that $\alpha_j \sim \frac{1}{X_j} < 1$. Multiplying out and neglecting terms with products yields

$$X_n - X_m \approx X_m \sum_{j=m+1}^{n} \alpha_j \sim X_m \left(n-m\right) E\left[\alpha\right]$$

and we arrive, with $a = E\left[\alpha\right]$ as shown in Section 4, at

$$X_n - X_m \approx a \left(n-m\right) X_m.$$

Figure 2 shows the age t of a story as a function of the number of users that dugg on the story s. In other words, each data point reflects the time at which a user diggs on the story s. The remarkable observation is that the number of digging users on stories, when stories are on the frontpage (time interval between two horizontal lines in Fig. 2), seem to be linear in time t. Figure 2 illustrates five stories with different eventual digg value. Hence, Figure 2 suggests that the user's increment $n-m$ for stories on the frontpage is proportional with time t, i.e. the n-th user and his appearance are related as

$$n = \beta t_n + \delta.$$

In that case (and assuming that also non-digging users arrive according to the linear law), we have approximately

$$X_n - X_m \approx a X_m \beta \left(t_n - t_m\right). \tag{14}$$

Figure 3 shows four five minute intervals in which the increment $f\left(x\right) = X_n - X_m$ is drawn versus the digg value $x = X_m$. When the time interval is short, hardly any linear correlation as suggested by (14) is observed.

When the time interval is longer, for example 1 h instead of 5 min, Figure 4 starts revealing the law in (14): the increments are proportional to the digg value at the beginning of the time interval. Hence, Figure 4 supports the claim that the digg values of a story are dependent and obey the law of proportionate effect, when the digg value is "sufficiently" high. The later condition is mathematically not precisely determined, because the law of proportionate

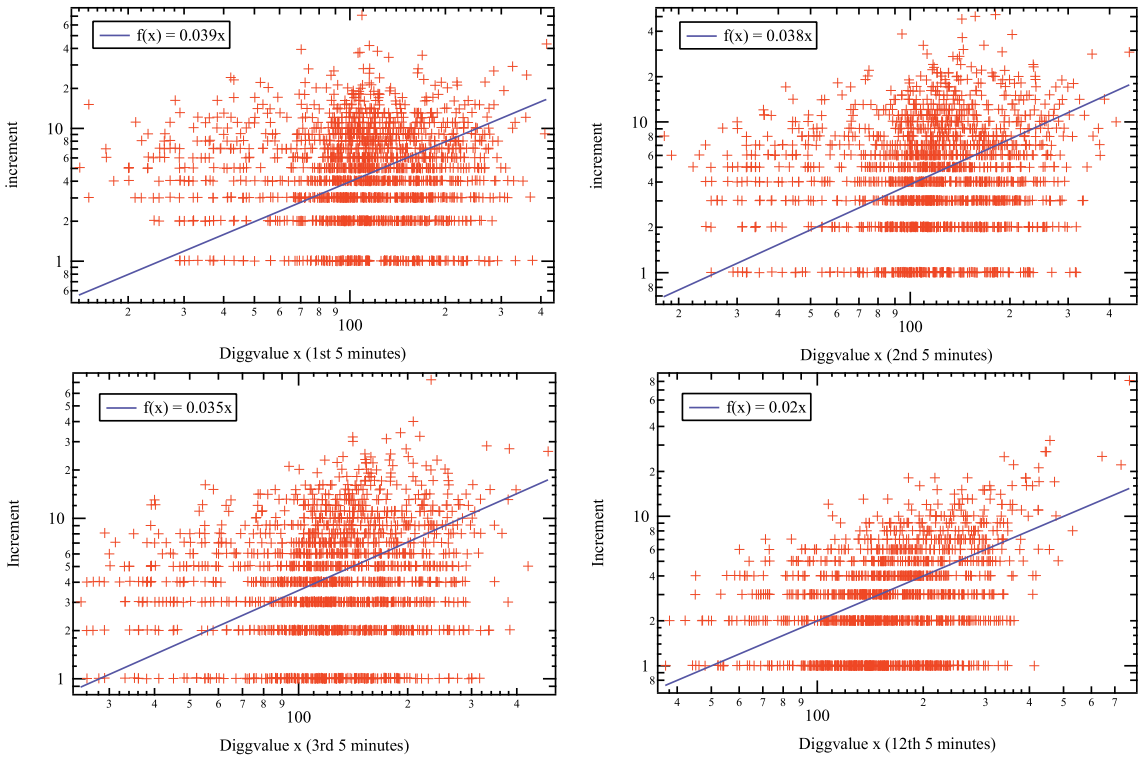

Fig. 3. (Color online) The increments during 5 min of stories on the frontpage versus their diggvalue at the beginning of the 5 min interval. Four 5 min intervals are shown and the linear fit of the increments $f(x)$ versus the digg value x. Notice that the linear fit $y = ax$ on a log-log scale is a line with slope 1 (or 45 degrees) and at $x = 1$ equal to $\log a$.

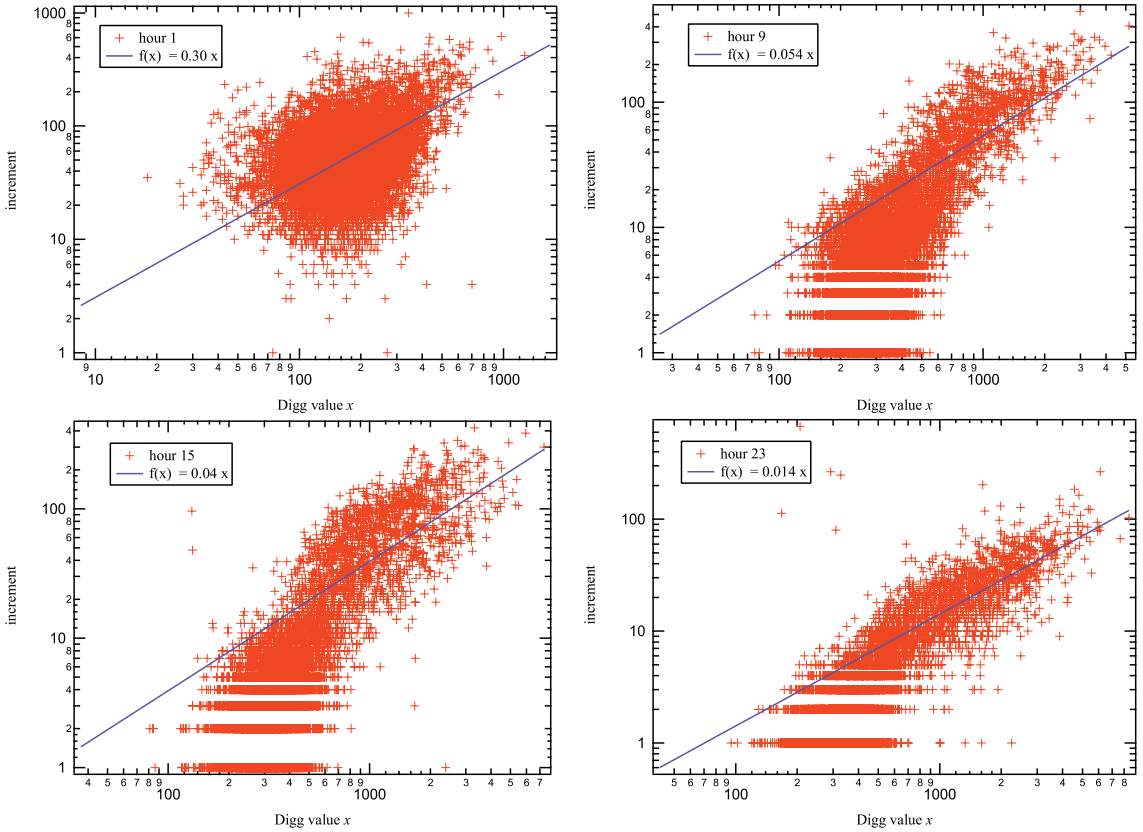

Fig. 4. (Color online) The increments of a story s versus its digg value x during 1 h. The only difference with Figure 3 is the duration of the interval (1 h versus 5 min).

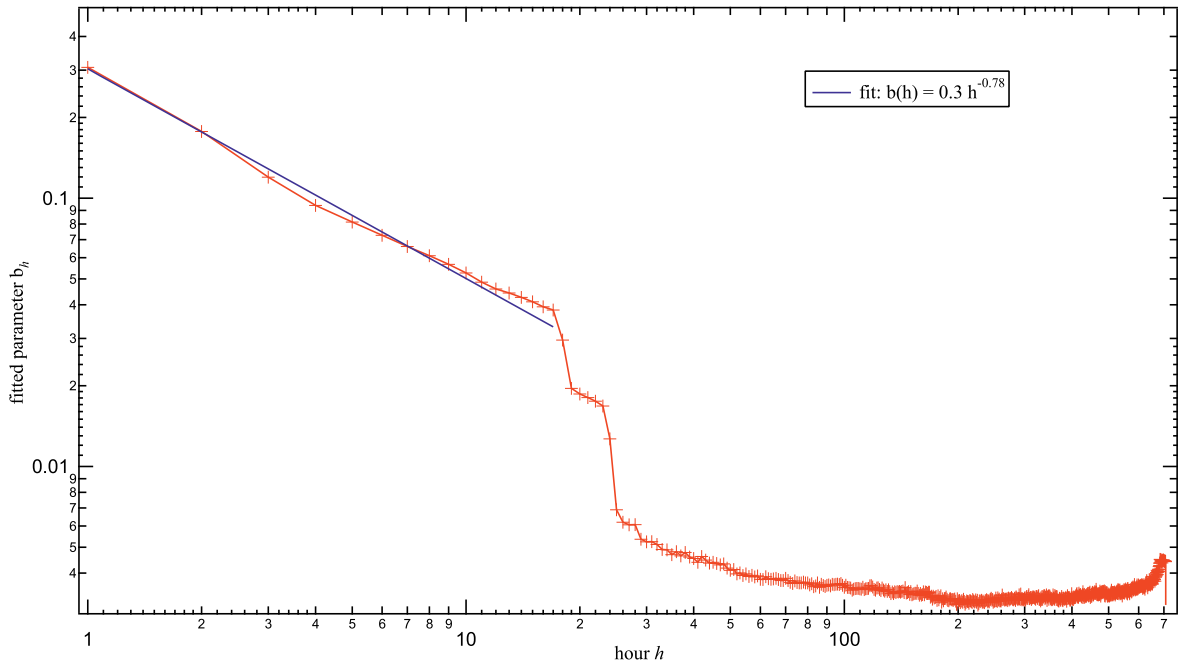

Fig. 5. (Color online) The slope, fitted from the increments during one hour versus the digg value, versus the sequential time intervals of 1 h. The fit of the first hours is also included.

effect is an asymptotic law (see Appendix A), yet observable in reality, i.e. when the digg values are finite. Figure 4 indicates that the tendency towards the asymptotic regime is rather slow. In view of the rather slow tendency towards the asymptotic regime, it is surprising that the distribution $\Pr[X_d = k]$ in Figure 1 is close to a lognormal. These observations question that the underlying process responsible for the nice lognormal on the front page is generated by the law of proportionate effect, because the time scales do not match. Rather, they suggest that the user population must be Gaussian-like distributed (as needed for finite digg values in Sect. 3.1) and that collective dependence, in a proportional fashion according to the hypothesis (12) constitutes the generating process.

Figure 5 shows the fits $a\beta(t_n - t_m)$ in (14), where $t_n - t_m = 1$ hour such that $a\beta \sim ct_n^{\gamma}$. Additionally approximating $\gamma \approx -1$ and using $n = \beta t_n + \delta$, we deduce that

$$a \sim \frac{c}{\beta t_n} \approx \frac{c}{n}$$

and this result (with $c \approx 0.3 < 1$) agrees with $a \leq \frac{1}{\mathcal{N}_s}$ derived in Section 3.2, but disagrees with the law of proportionate effect. For longer time intervals (exceeding 17 h), we observe in Figure 5 deviations in the fit of the slopes. For these long times, other effects (mainly due to ageing, the role of the user interface and the principles of human attention) start dominating.

6 Analysis of Wu and Huberman in [11]

After an initially fast increase, the digg value of a story flattens with time because the novelty of the story has passed. Wu and Huberman [11] have taken the ageing of a story into account. The effect of ageing means that the set $\{\alpha_k\}_{k\geq 1}$ in (A.1) starts decreasing after some time threshold k_c. Based on fitting experiments, Wu and Huberman [11] propose $\alpha_k = r_k b_k$, where $r_k = \exp\left(-0.4k^{0.4}\right)$ is the ageing factor and the set $\{b_k\}_{k\geq 1}$ of random variables is i.i.d. and with finite mean and variance. The Lindeberg conditions in [16] for the CLT state that $\sigma_k^2 = \mathrm{Var}[\alpha_k]$ should be small compared to $\sum_{j=1}^{\mathcal{N}_s} \sigma_k^2$ and the latter sum should tend to infinity when $\mathcal{N}_s \to \infty$. The decrease in α_k is so fast, that $\lim_{n\to\infty} \sum_{j=1}^{n} \sigma_k^2$ is finite, in which case the limiting distribution is not a Gaussian and, consequently, a lognormal cannot be explained from the law of proportionate effect! In that case, the function $g_k(x)$ in (4) cannot be a linear function. Nevertheless, Wu and Huberman [11] claim convergence to a Gaussian (lognormal) by referring to Embrechts and Maejima [17], Theorem 2, who show that $Z = \sum_{j=1}^{\infty} c_j(\lambda) X_j$ converges to a Gaussian with rate of the order of $O\left(\lambda^{-\alpha\beta}\right)$, where $\alpha > 0$ and where $\{X_j\}_{j\geq 0}$ is a set of i.i.d random variables with mean 0 and variance 1, and where $c_j(\lambda) = O\left(\lambda^{-\beta}\right)$ with $\beta > 0$. Yet, it is not clear whether this theorem applies to demonstrate convergence to a Gaussian (lognormal), because $r_j = \exp\left(-0.4j^{0.4}\right)$ is compared to $c_j(\lambda)$ and the r_j's depend on j, while the convergence in Embrechts and Maejima assumes that $c_j(\lambda) = O\left(\lambda^{-\beta}\right)$, for all j.

While we have shown in Section 5 that the law of proportionate effect cannot explain the experiments, the arguments of Wu and Huberman at least lack rigor, as sketched above. Apart from mathematical rigor, since ageing does not play a dominant role on the first frontpage,

their approach is essentially equal to the law of proportionate effect, which cannot explain the nice lognormal on the first frontpage.

7 Conclusion

We have investigated two different analyses that lead to a lognormal distribution, observed (see Fig. 1) for the digg value of a story, while on the first few frontpages. Usually, the lognormal distribution is the characteristic fingerprint of the law of proportionate effect. However, we found that the law of proportionate effect only seems to hold when a certain digg value of the story is reached, and not from the beginning of the digg counting. Moreover, the proportionality factor a is not constant as required by the law of proportionate effect, but dependent on the number of users as shown in our analysis of Section 3 and experimentially verified in Section 5. Finally, the governing difference equation (A.1) of the law of proportionate effect is not experimentally observed at times the story is on the first frontpage. Hence, the law of proportionate effect cannot explain the fast convergence towards the lognormal distribution of digg values on the first frontpage.

The second analysis just sums dependent Bernoulli random variables. We show that the dependence among users is reflected by (12),

$$\Pr\left[\text{user } j \text{ diggs on } s \middle| X_s(j) = m\right] = am.$$

This conditional probability (12) is a manifestation of proportionate effect and provides another way to verify proportional behavior. The conditional probability (12) illustrates how individual human behavior is affected by that of others, given that the individual can observe how the others react, e.g. via the digg value $X_s(j) = m$. Also, the above conditional probability shows that $a \leq \frac{1}{m}$ for each user m. At last, our analysis of Section 3 demonstrates that a lognormal distribution of the digg value D_s is obtained when the population $\mathcal{N}_s$ of potential diggers is normally distributed. Gaussian user populations are approximately measured[2].

At last, our analysis underlines the importance of the user population specifics, which are, by its asymptotic nature ($\mathcal{N}_s \to \infty$), not relevant in the law of proportionate effect.

We are grateful to Siyu Tang for her useful comments. Nobert Blenn is funded by TRANS (www.trans-research.nl).

Appendix A: The law of proportionate effect

As mentioned in [7] and in [18], Kapteyn considered in 1903 the equation

$$X_j - X_{j-1} = \alpha_j f(X_{j-1})$$

where the set $\{\alpha_j\}_{1\leq j\leq n}$ of random variables is mutually independent and identically distributed, equal to the distribution of the random variable α with mean $E[\alpha]$ and variance $\mathrm{Var}[\alpha]$. Moreover, the set $\{\alpha_j\}_{1\leq j\leq n}$ of random variables is also independent of the random variables $X_1, X_2, \ldots, X_n$. The special case where $f(x) = x$ reduces to

$$X_j = (1+\alpha_j) X_{j-1} \tag{A.1}$$

and the process that determines the sequence X_1, $X_2, \ldots, X_n$, given X_0, is said to obey the *law of proportionate effect*, which was first introduced by Gibrat [19].

After iterating the equation (A.1), we obtain

$$X_n = X_0 \prod_{j=1}^{n} (1+\alpha_j). \tag{A.2}$$

By the Central Limit Theorem [14] and assuming that any $\alpha_j > -1$, the sum $S_n = \sum_{j=1}^{n} \log(1+\alpha_j)$ of the i.i.d. random variables $\{\log(1+\alpha_j)\}_{j\geq 1}$, each with distribution identical to that of $\log(1+\alpha)$ with (finite) mean $E[\log(1+\alpha)] = \mu$ and variance $\sigma^2 = \mathrm{Var}[\log(1+\alpha)]$, converges to

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} N(0,1)$$

which implies that $S_n = \log(\prod_{j=1}^{n}(1+\alpha_j)) \xrightarrow{d} N(n\mu, n\sigma^2)$. Equivalently, $e^{S_n} = \prod_{j=1}^{n}(1+\alpha_j)$ tends, for large n, to a lognormal distribution with parameters $n\mu$ and $n\sigma^2$. Hence, we have shown that, for large n, X_n is asymptotically lognormally distributed with parameters $n\mu$ and $n\sigma^2$, that are linear in n.

There is a continuous variant of the law of proportionate effect. In biology, the growth in the number $n(t)$ of items of a same species over time t can be modelled by the following first order differential equation

$$\frac{dn(t)}{dt} = r(t)\, n(t).$$

which relates the growth (change in the population) as proportional to the population $n(t)$ and the proportionality factor $r(t)$ is time dependent. The general solution is, for $t > a$,

$$\log n(t) = \log n(a) + \int_a^t r(u)\, du.$$

When we additionally assume that $r(t)$ changes at some times $a = t_0 < t_1 < t_2 < \ldots < t_m = t$, where t_j are random time moments, then

$$\int_a^t r(u)\, du = \sum_{j=1}^{m} \int_{t_{j-1}}^{t_j} r(u)\, du = \sum_{j=1}^{m} R_j$$

where

$$R_j = \int_{t_{j-1}}^{t_j} r(u)\, du = r(\xi_j)(t_j - t_{j-1}) \text{ and } \xi_j \in [t_{j-1}, t_j]$$

[2] See e.g. http://www.alexa.com, *Alexa the Web Information Company*, Alexa Internet, Inc., 2010.

$$\rho = \frac{\{\Pr[\text{user } k \text{ diggs on } s | \text{ user } l \text{ diggs on } s] - \Pr[\text{user } k \text{ diggs on } s]\} \Pr[\text{user } l \text{ diggs on } s]}{\sqrt{\Pr[\text{user } k \text{ diggs on } s](1 - \Pr[\text{user } k \text{ diggs on } s]) \Pr[\text{user } l \text{ diggs on } s](1 - \Pr[\text{user } l \text{ diggs on } s])}}.$$

is a random variable with mean $\mu_j = E[r(\xi_j)(t_j - t_{j-1})]$. Assuming that the Central Limit Theorem can be applied, the set of random variables $\{R_j\}_{1 \le j \le m}$ tends to a Gaussian $N(m\mu, m\sigma^2)$, and the lognormal distribution of $n(t)$ for large t then follows in the usual way. Again, the mean is linear in the time t because $\frac{t-a}{m} = E[\Delta t]$, the average time-spacing.

Appendix B: The variance of $X_s(j)$

In order to compute the variance $\mathrm{Var}[X_s(j)] = E\left[X_s^2(j)\right] - (E[X_s(j)])^2$, we first rewrite $X_s^2(j)$ using the definition (3) as

$$\begin{aligned} X_s^2(j) &= \sum_{k=1}^{j-1}\sum_{l=1}^{j-1} 1_{\{\text{user } k \text{ diggs on } s\}} 1_{\{\text{user } l \text{ diggs on } s\}} \\ &= \sum_{k=1}^{j-1} 1_{\{\text{user } k \text{ diggs on } s\}} \\ &\quad + 2\sum_{k=1}^{j-1}\sum_{l=1}^{k-1} 1_{\{\text{user } k \text{ diggs on } s\}} 1_{\{\text{user } l \text{ diggs on } s\}} \end{aligned}$$

and

$$\begin{aligned} X_s^2(j) &= X_s(j) \\ &\quad + 2\sum_{k=1}^{j-1}\sum_{l=1}^{k-1} 1_{\{(\text{user } k \text{ diggs on } s) \cap (\text{user } l \text{ diggs on } s)\}}. \end{aligned}$$

Taking the expectation gives

$$\begin{aligned} E\left[X_s^2(j)\right] &= E[X_s(j)] \\ &+ 2\sum_{k=1}^{j-1}\sum_{l=1}^{k-1} \Pr[(\text{user } k \text{ diggs on } s) \cap (\text{user } l \text{ diggs on } s)] \end{aligned}$$

where user l proceeds user k. Invoking the conditional probability, finally results in

$$\begin{aligned} E\left[X_s^2(j)\right] &= E[X_s(j)] \\ &+ 2\sum_{k=1}^{j-1}\sum_{l=1}^{k-1} \Pr[\text{user } k \text{ diggs on } s | \text{ user } l \text{ diggs on } s] \\ &\quad \times \Pr[\text{user } l \text{ diggs on } s]. \quad \text{(B.1)} \end{aligned}$$

In general, we can upper bound $E\left[X_s^2(j)\right]$ as

$$\begin{aligned} E\left[X_s^2(j)\right] &\le E[X_s(j)] + 2\sum_{k=1}^{j-1}\sum_{l=1}^{k-1} \Pr[\text{user } l \text{ diggs on } s] \\ &= E[X_s(j)] + 2\sum_{k=1}^{j-1} E[X_s(k)] \le (j-1)^2 \end{aligned}$$

from which we obtain an upper bound of the variance

$$\mathrm{Var}[X_s(j)] \le E[X_s(j)] + 2\sum_{k=1}^{j-1} E[X_s(k)] - (E[X_s(j)])^2.$$

If users were to digg independently so that

$$\Pr[\text{user } k \text{ diggs on } s | \text{ user } l \text{ diggs on } s] = \Pr[\text{user } k \text{ diggs on } s]$$

then the conditional probability in (B.1) would reduce to

$$\begin{aligned} E\left[X_s^2(j)\right] &= E[X_s(j)] \\ &\quad + 2\sum_{k=1}^{j-1} \Pr[\text{user } k \text{ diggs on } s] E[X_s(k)]. \end{aligned}$$

In Digg, users are positively correlated[3] such that

$$\Pr[\text{user } k \text{ diggs on } s | \text{ user } l \text{ diggs on } s] \ge \Pr[\text{user } k \text{ diggs on } s]$$

which leads to a higher second moment $E\left[X_s^2(j)\right]$ in (B.1) than for independent users. However, the mean $E[X_s(j)] \le j-1$ in Digg is also larger. Since $\mathrm{Var}[X_s(j)] \le (j-1)^2 - (E[X_s(j)])^2$, the mean decreases the variance quadratically, which may lead to a relatively small variance. In other words, the stronger the proportionate effect (measured via the proportionality factor a in (12)), the smaller the variance $\mathrm{Var}[X_s(j)]$ and the better the mean $E[X_s(j)]$ approximates the random variable $X_s(j)$.

References

1. Rob McGann, *Internet Edges Out Family Time More Than TV Time* (ClickZ.com, 2005)
2. C. Nuttall and D. Gelles, *Facebook becomes bigger hit than Google* (Financial Times, 2010)

[3] Applying the definition [14], p. 30 of the linear correlation coefficient ρ yields

see equation above.

3. G. Szabo, B.A. Huberman, Commun. ACM **53**, 80 (2010)
4. C. Castellano, S. Fortunato, V. Loreto, Revi. Mod. Phys. **81**, 591 (2009)
5. K. Lerman, A. Galstyan, *Proceedings of the First Workshop on online Social Networks (WOSP '08)* (ACM, New York, 2008), pp. 7–12
6. S. Tang, N. Blenn, C. Doerr, P. Van Mieghem, to appear in IEEE Transactions on Multimedia
7. E.L. Crow, K. Shimizu, *Lognormal distributions, Theory and Applications* (Marcel Dekker, Inc. New York, 1988)
8. E. Limpert, W.A. Stahel, M. Abbt, Bioscience **51**, 341 (2001)
9. W. Shockley, Proceedings of the IRE **45**, 279 (1957)
10. F. Radicchi, S. Fortunato, C. Castellano, Proc. Natl. Acad. Sci. USA **105**, 17268 (2008)
11. F. Wu, B.A. Huberman, Proc. Natl. Acad. Sci. USA **104**, 17599 (2007)
12. C. Doerr, S. Tang, N. Blenn, P. Van Mieghem, *Are Friends Overrated? A Study of the Social News Aggregator Digg.com* (IFIP Networking, 2011)
13. R.K. Pan, S. Sinha, New J. Phys. **12**, 115004 (2010)
14. P. Van Mieghem. *Performance Analysis of Communications Systems and Networks* (Cambridge University Press, Cambridge, U.K., 2006)
15. Deese, Kaufman, J. Exp. Psychol. **54**, 180 (1957)
16. W. Feller, *An Introduction to Probability Theory and Its Applications*, 2nd edn. (John Wiley & Sons, New York, 1971), Vol. 2
17. P. Embrechts, M. Maejima, Zeitung für Wahrscheinlichkeitstheorie und verwandte Gebiete **68**, 191 (1984)
18. M. Armate, Mathématiques et sciences humaines **129**, 5 (1995)
19. R. Gibrat, Bulletin de la Statistique Générale de la France **19**, 469 (1930)