Correspondence

Human Psychology of Common Appraisal: The Reddit Score

Piet Van Mieghem

***Abstract*—The Reddit score reflects a common appraisal by a community of Reddit subscribers of a submitted item, called a story. The general random walk with random maximum boundary is demonstrated to describe the distribution function of the Reddit score of an arbitrary story in the online social news aggregator Reddit.com. Exponential tails, predicted by the analysis, are observed, while a curious intermediate "power law-like" region seems to correspond to a remarkable empirical observation that the total number of downvotes depends in "power law" fashion on the total number of upvotes. Stronger even, those downvotes increase faster than the upvotes, which is a surprising fact that asks for a (socio-psychological?) explanation.**

***Index Terms*—Collective behavior, network, random walk, voting.**

I. Introduction

Human appraisal in online social news aggregators such as Digg.com and Reddit.com is reflected by the popularity counter, respectively, the Digg value or the Reddit score, assigned to a submitted story.[1] The popularity counter of a story shows the number of net votes that users can give. In Digg.com, which is studied in [2], a user can increase the Digg value by one, while in Reddit.com, a user can either increase or decrease the Reddit score by one. The higher the popularity counter, the more votes a story has received and the more attractive the story is, which often results in a better position on the webpage. Highly popular stories create impact on a large population. Moreover, the popularity score is a collective appraisal: individual users are biased in their voting behavior by the current score they observe. This "open" voting process is thus entirely different in nature than "closed" voting, where users do not see a score. Knowing how stories get popular and understanding the voting dynamics in "cyberspace" constitutes a major area of interest, especially in direct marketing and all types of open elections.

Here we present an accurate model, based on the general random walk, for the distribution function of the Reddit score of an arbitrary story in Reddit.com. We show that, depending on a few *sensitive* parameters, various observed laws, from lognormal to power laws, can be imitated, which underlines that caution is in order, certainly in an era where power laws are seen everywhere [1]. Comparison with Reddit experiments from September 2010 to November 2010 on over a million stories verifies the qualitative behavior of the model. The experiments also exhibit a rather strange and counter-intuitive law for the downvoting of stories. While upvoting of stories is generally described by preferential attachment or the law of proportionate effect [4], [6], downvoting seems more complex. We found that downvoting is related to upvoting in a power law fashion, but the underlying (socio-psychological?) law that generates this observed power law dependence is unknown, at least to us.

Manuscript received February 17, 2011; revised June 06, 2011; accepted July 21, 2011. Date of current version November 18, 2011. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yiannis Andreopoulos.

The author is with the Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, 2600 GA Delft, The Netherlands (e-mail: P.F.A.VanMieghem@tudelft.nl).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2011.2165054

[1]Any submitted digital, multi-media type of data such as news, comments, and pictures with a separate entry and popularity counter is here called a story.

II. Reddit Score of a Story

We model the Reddit score X_k of a story at discrete time k as a general random walk [5, p. 202]. The state j in the general random walk corresponds to a Reddit score equal to j. Each transition in the general random walk is caused by a user $k+1$, who observes the score $X_k = j$ and either increases the score with probability $p_j = \Pr[X_{k+1} = j+1 | X_k = j]$, decreases the score with probability $q_j = \Pr[X_{k+1} = j-1 | X_k = j]$ or does not change the score with probability $r_j = \Pr[X_{k+1} = j | X_k = j]$. Clearly, $p_j + q_j + r_j = 1$. The steady-state of the general random walk over N possible states corresponds to the eventual Reddit score of the story and is given by [5, p. 207]

$$\Pr[X_\infty = j | N] = \frac{\prod_{m=0}^{j-1} \frac{p_m}{q_{m+1}}}{1 + \sum_{k=1}^{N} \prod_{m=0}^{k-1} \frac{p_m}{q_{m+1}}} \tag{1}$$

where $0 \le j \le N$ and $\prod_{m=a}^{b} f(m) = 1$ if $a > b$. Similar to the Digg analysis [4], we have tacitly assumed that N is a given constant. The largest possible Reddit score N equals the total number of people that can change the story's score, because there is always a nonzero probability that all those users vote positively on the story. In reality, however, the total number of users N for an arbitrary story is a random variable with certain distribution $\Pr[N = k]$ and maximum n. The number N of users that can vote on a story is also dependent on the popularity of the story, because, in Reddit, stories are ranked corresponding to their Reddit score. This means that less popular stories fade away to uninteresting pages that are less often visited than highly popular stories. In short, the number N of users is quite a complicated random variable, that is weakly dependent on X_k. In the sequel, however, we assume that N is independent of X_∞, to allow analytic modeling.

The usual way to incorporate the randomness of N is to apply the law of total probability [5], yielding

$$\begin{aligned}\Pr[X_\infty = j] &= \sum_{l=0}^{n} \Pr[X_\infty = j | N = l] \Pr[N = l] \\ &= \sum_{l=0}^{n} \frac{\prod_{m=0}^{j-1} \frac{p_m}{q_{m+1}} 1_{\{1 \le j \le l\}}}{1 + \sum_{k=1}^{l} \prod_{m=0}^{k-1} \frac{p_m}{q_{m+1}}} \Pr[N = l]\end{aligned}$$

and we arrive at the Reddit score probability density function (pdf)

$$\Pr[X_\infty = j] = \prod_{m=0}^{j-1} \frac{p_m}{q_{m+1}} \sum_{l=j}^{n} \frac{\Pr[N = l]}{1 + \sum_{k=1}^{l} \prod_{m=0}^{k-1} \frac{p_m}{q_{m+1}}}. \tag{2}$$

The aim of this letter lies in explaining (2) in Section II-A and verifying (2) with empirical data in Section II-B by estimating the ratio p_m/q_{m+1} and the total population distribution $\Pr[N = l]$.

A. General Observations on (2)

From (2), we deduce the general difference equation for the Reddit score pdf

$$y_j - \frac{q_{j+1}}{p_j} y_{j+1} = \frac{\left(\prod_{m=0}^{j-1} \frac{p_m}{q_{m+1}}\right) \Pr[N=j]}{\left(1 + \sum_{k=1}^{j} \prod_{m=0}^{k-1} \frac{p_m}{q_{m+1}}\right)} \tag{3}$$

where $y_j = \Pr[X_\infty = j]$. After rewriting (3), we obtain

$$\frac{\Pr[X_\infty = j+1]}{\Pr[X_\infty = j]} = \frac{p_j}{q_{j+1}} \left(1 - \frac{\prod_{m=0}^{j-1} \frac{p_m}{q_{m+1}}}{1 + \sum_{k=1}^{j} \prod_{m=0}^{k-1} \frac{p_m}{q_{m+1}}} \frac{\Pr[N=j]}{\Pr[X_\infty = j]}\right) \tag{4}$$

which shows that $\Pr[X_\infty = j]$ is increasing (decreasing) in j provided the right-hand side of (4) is larger (smaller) than 1. Since the term in brackets lies in the interval $[0, 1]$, the factor p_j/q_{j+1} determines the increase or decrease of $\Pr[X_\infty = j]$ in j.

Usually, the population has a bell-shape distribution around its mean $E[N] = \mu$ with exponentially decaying tails. Thus, for Reddit scores $j < \mu - x\sigma$ or $j > \mu + x\sigma$, where the standard deviation $\sigma = \sqrt{\mathrm{Var}[N]}$ and x is about 3 or more (for a Gaussian), the right-hand side of (3) is vanishingly small, such that

$$\frac{\Pr[X_\infty = j+1]}{\Pr[X_\infty = j]} \approx \frac{p_j}{q_{j+1}}.$$

Thus, $\Pr[X_\infty = j]$ is increasing (decreasing) in j if $p_j/q_{j+1} > 1$ ($p_j/q_{j+1} < 1$) for all j in the interval $0 \le j < \mu - x\sigma$. In the interval $j > \mu + x\sigma$ and assuming a monotonous law for p_j/q_{j+1} (also inspired by Fig. 2 below), $\Pr[X_\infty = j]$ can only be decreasing in j, because $\Pr[X_\infty = n+1] = 0$ and $\Pr[X_\infty = n] > 0$, where n is the maximum possible value of X_∞. In between these two regimes, $\mu - x\sigma < j < \mu + x\sigma$, there is a transition regime, where the right-hand side of (3) achieves a maximum. Generally, the largest change in $\Pr[X_\infty = j]$ is expected in this transition regime. Notice that $\mu - x\sigma^2$ can be negative, that $\mu + x\sigma^2$ can exceed n, and that only the transition regime can be observed in reality.

Since $\sum_{k=1}^{l} \prod_{m=0}^{k-1} (p_m/q_{m+1})$ increases with l, we can upper bound (2) as

$$\frac{\left(\prod_{m=0}^{j-1} \frac{p_m}{q_{m+1}}\right) \Pr[N=j]}{1 + \sum_{k=1}^{j} \prod_{m=0}^{k-1} \frac{p_m}{q_{m+1}}} \le y_j \le \frac{\left(\prod_{m=0}^{j-1} \frac{p_m}{q_{m+1}}\right) \Pr[N \ge j]}{1 + \sum_{k=1}^{j} \prod_{m=0}^{k-1} \frac{p_m}{q_{m+1}}}.$$

In the tail region where $\Pr[N = j] \lesssim \Pr[N \ge j]$, both lower and upper bound tend to each other, which reveals that $\Pr[N = j]$ and $\prod_{m=0}^{j-1} p_m/q_{m+1}$ determine the distribution of the Reddit score for large j. Our analysis here in Section II-A underlines the importance of the population distribution $\Pr[N = k]$. In most studies (e.g., [6]), which rely on asymptotic laws such as the law of proportionate effect, the population distribution $\Pr[N = k]$ is ignored so far.

B. Experiments

We study the Reddit scores of 1 740 066 stories. Fig. 1 shows the pdf $\Pr[X_\infty = k]$ and illustrates that the tail region is exponentially decaying, whereas an intermediate region (see insert) is observed that resembles a "power law", which is conceivably related to a remarkable observation shown in Fig. 2.

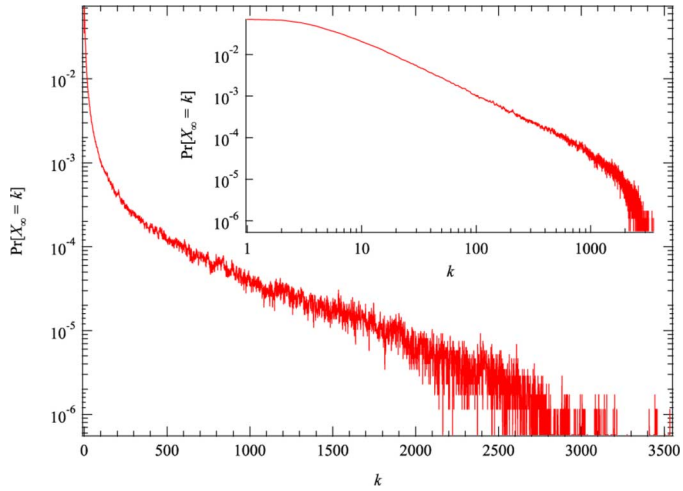

Fig. 1. The pdf of the Reddit score on a log-lin and log-log scale. The tail region is fitted as $\Pr[X_\infty = k] \simeq 3.10^{-4} \exp(-k/511)$, whereas the intermediate "power law" region is fitted by $\Pr[X_\infty = k] \simeq 0.62 k^{-1.38}$.

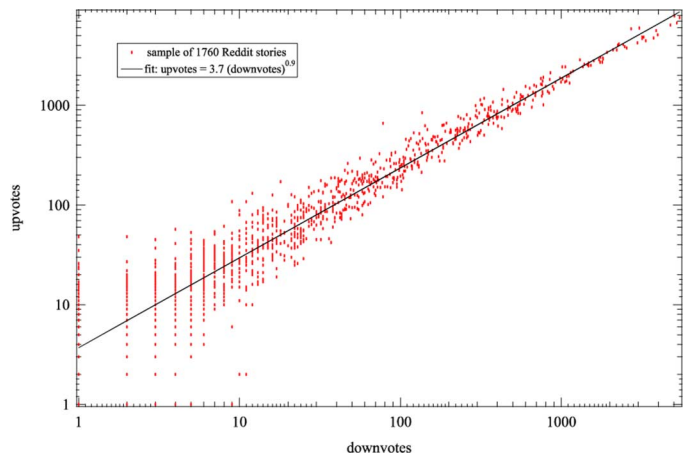

Fig. 2. Scatter plot of the pair (upvotes, downvotes) for a sample of 1760 Reddit stories out of 1 740 066 stories.

Fig. 2 draws the up- versus downvotes for 1760 stories[2] and reveals the power law relation between up- and downvotes of a story. The Reddit score is $X_\infty = U_\infty - D_\infty$, where U_∞ and D_∞ are the eventual number of upvotes and downvotes of a random story. The fit of all 1 740 066 stories is $U_\infty = cD_\infty^\beta$, where $\beta \approx 0.894$ and $c \approx 3.80$. Since $X_\infty \ge 0$, there must hold that $cD_\infty^\beta - D_\infty \ge 0$, or $c \ge D_\infty^{1-\beta}$, and hence, $c \ge (\max D_\infty)^{1-\beta}$, a condition that is met. We may assume that the promotion probability $p_j = a(j+1)$ for $j \ge 0$ is due to the law of proportionate effect, as explained for the online social news aggregator Digg in [4]: when a user observes a score j, the probability that he will upvote the story's score (i.e., increase by 1) is proportional to j. This proportionate upvoting gives rise to a lognormal distribution. While the law of proportionate effect [4] for upvotes can be intuitively understood, Fig. 2 is surprising: also the downvoting of stories seems positively correlated to the score j that a user sees. Intuitively, one might have expected that a story with a high score j would receive downvotes inversely proportional to its score, i.e., $q_j \sim 1/j$. However, experiments point to the reverse: also a high score j will suffer from high downvotes. Fig. 2 seems to reflect the saying "Tall trees capture much wind". Popularity seems to have its adversaries. Very recently,

[2]Since the entire file of downvotes and upvotes was too large to draw in LaTex (the $*$.eps file is 65 Mb), a random sample has been taken, which, visually, represents the entire set very well.

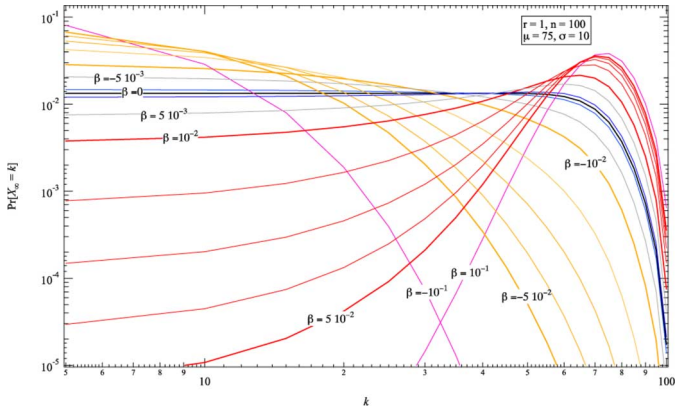

Fig. 3. Analytic computations of pdf of the Reddit score (5) for various β, $r = 1$, $n = 100$ and a (discrete) Gaussian population $\Pr[N = k] = (1/\sigma\sqrt{2\pi})\exp(-(k-\mu)^2/2\sigma^2)$ with $\mu = 75$ and $\sigma = 10$.

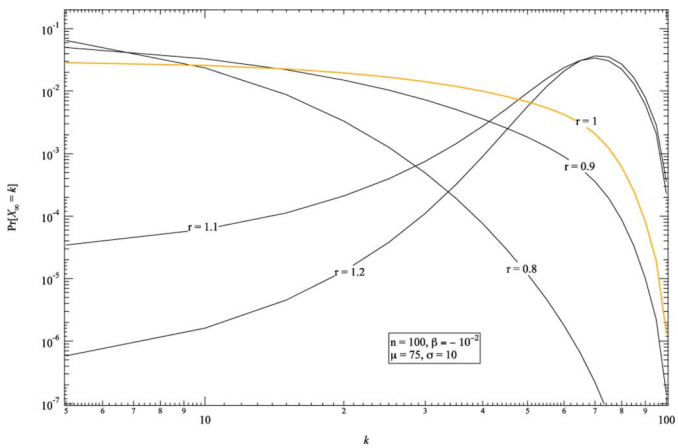

Fig. 4. Same setting as in Fig. 3, except that r is changed, while $\beta = -10^{-2}$.

Thelwall *et al.* [3] have observed in a Twitter sentimental analysis that important events in Twitter are associated with increases in average negative sentiment strength. While also positive sentiment can lead to popular events in Twitter, negative sentiment seems to be more central. Hence, our finding about the proportionate effect for downvotes is in line with Thelwall's sentimental analysis: positive events are likely to trigger negative emotions in social networks.

Assuming that $p_j = a(j+1)$ for $j \geq 0$, due to the law of proportionate effect, then the fit in Fig. 2 may suggest that the downgrading probability $q_j \sim j^{1/0.9} = j^{1.11}$ and that the important fraction $p_j/q_{j+1} \sim (j+1)^{\beta}$, where $\beta \approx -0.1$. The surprising fact is that p_j/q_{j+1} decreases with j. The analysis of Section II-A concludes that, for $(p_j/q_{j+1}) < 1$, the probability $\Pr[X_\infty = j]$ is decreasing in j, which agrees with Fig. 1. The downvoting law $D_\infty = (U_\infty/c)^{1/\beta}$, where $1/\beta > 1$ and implying that D_∞ increases faster than U_∞, is a strange empirical observation, which asks for a deeper explanation.

III. Reddit Pdf With $p_m/q_{m+1} = r(m+1)^\beta$

Inspired by the empirical observation in Fig. 2, we choose $p_m/q_{m+1} = r(m+1)^\beta$, such that $\prod_{m=0}^{j-1}(p_m/q_{m+1}) = r^j(j!)^\beta$ and (1) becomes

$$\Pr[X_\infty = j] = r^j(j!)^\beta \sum_{l=j}^{n} \frac{\Pr[N=l]}{\sum_{k=0}^{l} r^k (k!)^\beta}. \tag{5}$$

Computations of (5) for n above 1000 are hard due to the factorial $k!$ so that we did not succeed to fit the empirical data in Fig. 1 with (5). Since smaller n sketch the behavior well, we illustrate (5) for $n = 100$. Figs. 3 and 4 demonstrate the sensitivity of $\Pr[X_\infty = j]$ on β and r, respectively, for $n = 100$. The curves for $\beta = -10^{-2}$ and $\beta = -5.10^{-3}$ in Fig. 3 exhibit a "seemingly" linear decreasing intermediate region on a log-log scale, that is also observed in Fig. 1. The sensitivity of changes in the mean and variance of the Gaussian population $\Pr[N = l]$ were found to be small.

There is no explicit, analytic form for $\sum_{k=0}^{l} r^k (k!)^\beta$, unless $\beta = 0$. For $\beta = 0$, (5) is bounded, for large j and supposing that $r > 1$, by

$$\begin{aligned}\Pr[X_\infty = j] &< r^j(r-1) \max_{j \leq l \leq n} \Pr[N=l] \sum_{l=j}^{n} \frac{1}{r^{l+1}-1} \\ &\approx r^j(r-1) \max_{j \leq l \leq n} \Pr[N=l] \frac{1}{r^j} \frac{1 - \frac{1}{r^{n-j+1}}}{r-1}.\end{aligned}$$

Thus

$$\Pr[X_\infty = j] \lessapprox \max_{j \leq l \leq n} \Pr[N=l] \left(1 - \frac{r^j}{r^{n+1}}\right)$$

which implies that, for large j, $\Pr[X_\infty = j]$ decreases at least exponentially fast. The second factor in the upper bound is a cut-off function: when j is small, it is approximately 1, but above $j > ((n+1)\log r - \log 2/\log r)$, it decreases exponentially fast to zero (see $\beta = 0$ curve in Fig. 3).

IV. Conclusion

The Reddit score is shown to be well modeled by a general random walk with probabilistic boundary N equal to the total number of users that are able to vote on the story. The general random walk with the law for $p_m/q_{m+1} = r(m+1)^\beta$ closely models empirical data (Fig. 1), while the analysis points to the very high sensitivity of r and β: small changes cause a largely different behavior (Fig. 3). The influence of the Gaussian population parameters via $\Pr[N = k]$ on $\Pr[X_\infty = j]$ are found to be much less pronounced. Moreover, both exponential and power law regions appear and demonstrate that often made claims about lognormal or power-law behavior in empirical data need to be interpreted with caution.

The Reddit score distribution in Fig. 1 exhibits an exponentially decreasing tail (predicted by the analysis) and possesses a "power law-like" intermediate region. The origin of the power law is triggering, but we speculate that it is due to an observed power law dependence $U_\infty = cD_\infty^\beta$ between the eventual up- and downvotes of stories (Fig. 2). While we are unable to explain $U_\infty = cD_\infty^\beta$, we are looking forward to receiving clarifying arguments.

Acknowledgment

The author would like to thank N. Blenn for the data.

References

[1] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Rev.*, vol. 51, no. 4, pp. 661–703, 2009.
[2] S. Tang, N. Blenn, C. Doerr, and P. Van Mieghem, "Digging in the Digg online social network," *IEEE Trans. Multimedia*, vol. 13, no. 5, pp. 1163–1175, Oct. 2011.
[3] M. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment in Twitter events," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 62, no. 2, pp. 406–418, 2011.
[4] P. Van Mieghem, N. Blenn, and C. Doerr, "Lognormal distribution in the Digg online social network," *Eur. Phys. J. B*.
[5] P. Van Mieghem, *Performance Analysis of Communications Systems and Networks*. Cambridge, U.K.: Cambridge Univ. Press, 2006.
[6] F. Wu and B. A. Huberman, "Novelty and collective attention," *Proc. Nat. Acad. Sci.*, vol. 104, no. 45, pp. 17599–17601, Nov. 2007.