Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media

Social Dynamics of Digg

Tad Hogg
Institute for Molecular Manufacturing
Palo Alto, CA 94301, USA

Kristina Lerman
USC Information Sciences Institute
Marina del Rey, CA 90292, USA

Abstract

Online social media often highlight content that is highly rated by neighbors in a social network. For the news aggregator Digg, we use a stochastic model to distinguish the effect of the increased visibility from the network from how interesting content is to users. We find a wide range of interest, and distinguish stories primarily of interest to users in the network from those of more general interest to the user community. This distinction helps predict a story's eventual popularity from users' early reactions to the story.

Introduction

The Social Web enables people to connect, create and organize information. Stochastic models can help evaluate this technology, e.g., for the social news aggregator Digg (Lerman 2007; Hogg and Lerman 2009). The resulting model did not address a key aspect of social media: the extent to which links indicate commonality of user interests. This paper addresses this limitation with an extended model. The new model also accounts for the daily variation in user activity and the variation in votes a story receives before it is promoted, which the prior model ignored.

Digg users submit and rate news stories by voting on them. Digg promotes a few percent of submitted stories to the highly visible *front page* based on the user votes. Submitted stories appear in the *upcoming* stories list, where they remain for 24 hours or until promoted to the front page. Both upcoming and front page lists show stories in reverse chronological order, with 15 stories to a page. Digg allows users to designate friends. When user A lists user B as a *friend*, A can watch the activities of B but not vice versa. We call A the *fan* of B.

Fig. 1 illustrates how a story accumulates votes. The final number of votes varies widely among the stories. Some promoted stories accumulate thousands of votes, while others muster only a few hundred. A challenge for understanding this variation is the interaction between how Digg displays stories (their *visibility*) and their *interestingness* to users. Models accounting for the structure of the Digg interface can help separate these interactions.

We collected data from May and June 2006. The May data set has stories submitted to Digg May 25-27, 2006. By periodically scraping Digg, we collected at least 4 observations for each of 2152 stories, submitted by 1212 distinct users. Of these stories, 510, by 239 distinct users, were promoted. We followed the promoted stories over a period of several days. This May data set also records the location of the stories on the upcoming and front pages as a function of time. The June data set has 201 stories promoted between June 27 and 30, 2006, with the names of the first 216 voters on each story. In October 2009, we used the Digg API to obtain the time of each vote, the final number of votes the story received, and the time of promotion.

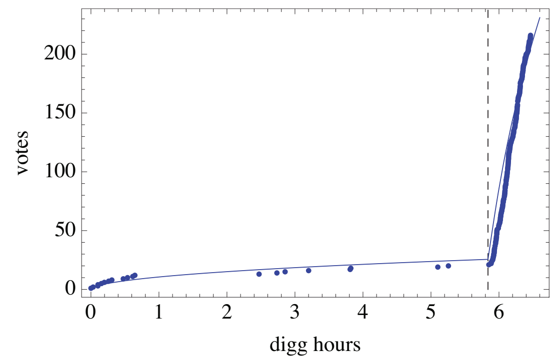

Figure 1: The number of votes vs. time, in Digg hours, for a promoted story in June 2006. The curve is the model solution and the dashed vertical line shows when the story was promoted. This story eventually received 2566 votes.

We extracted the social network of the top-ranked 1020 Digg users as of June 2006. Since the original network did not contain all the voters in our data, we augmented it in February 2008 with friends of about $15,000$ additional users. Many of these users added friends between June 2006 and February 2008. Although Digg does not provide the time of link creation, it lists the links in reverse chronological order and gives the date the friend joined Digg. By eliminating friends who joined Digg after June 30, 2006, we were able to reconstruct the links for all voters in our data. This data allows us to identify, for each vote, whether the user was a fan of any prior voter on that story, in which case the story would have appeared in the friends interface for that user. Votes by fans account for 6% of the votes in the June data set and about 3% of the front page votes.

Social Dynamics of Digg

We extend the previous stochastic model (Hogg and Lerman 2009) to distinguish votes from fans and non-fans. We take the number of users to be constant over our short sample period. Activity on Digg varies considerably over the course of a day, so we define the "digg time" between two events as the number of votes on front page stories between those events (Szabo and Huberman 2008). We scale the measure by defining a "Digg hour" to be the average number of front page votes in an hour, i.e., 2500 for our data set. This gives the average rates of growth for votes from fans and non-fans of prior voters, v_F and v_N, respectively as:

$$\frac{dv_F}{dt} = \omega r_F P_F F \quad \text{and} \quad \frac{dv_N}{dt} = \omega r_N P_N N$$

where t is the Digg time since the story's submission and ω is the average rate a user visits Digg. v_N includes the story's submitter. P_F and P_N denote the story's *visibility* and r_F and r_N denote the story's *interestingness* to users who are fans or not of prior voters, respectively. Visibility depends on the story's state (e.g., whether it has been promoted), as discussed below. Interestingness is the probability a user who sees the story will vote on it. Nominally people become fans of those whose contributions they consider interesting, suggesting fans likely have higher interest in stories.

These voting rates depend on F (N), the numbers of users who have not yet seen the story and who are (are not) fans of prior voters. The quantities change according to

$$\begin{aligned}\frac{dF}{dt} &= -\omega P_F F + \rho N \frac{dv}{dt}\\ \frac{dN}{dt} &= -\omega P_N N - \rho N \frac{dv}{dt}\end{aligned}$$

where $v = v_F + v_N$ and ρ is the probability a user who has not yet seen the story and is not a fan of a prior voter is a fan of the most recent voter. The first term in each equation is the rate the users see the story. Table 1 lists the model parameters, estimated as described below.

Initially, the story has one vote (from the submitter) and the submitter has S fans, so $v_F(0) = 0$, $v_N(0) = 1$, $F = S$ and $N = U - S - 1$ where U is the number of active users. Over time, a story becomes less visible as it moves down the upcoming or (if promoted) front page lists.

We assume a fan easily sees the story via the friends interface, so $P_F = 1$. Other users must find the story so P_N depends on how users navigate through the upcoming or front page lists, i.e., the "law of surfing" (Huberman et al. 1998) which gives an inverse Gaussian distribution of the number of pages a user visits, with mean μ and variance μ^3/λ.

The page number of a story on the upcoming page q and the front page p at time t is $p = k_{\text{f}}(t - T_{\text{promotion}}) + 1$ and $q = k_{\text{u}} t + 1$ where $T_{\text{promotion}}$ is the story's promotion time. Upcoming stories are less popular than the front page, modeled by a fraction $c < 1$ of visitors viewing the upcoming stories. Combining these effects gives a model of visibility (Hogg and Lerman 2009) which determines P_N.

Promotion to the front page depends on the number of votes, which we model by the probability $P(v)$ a story is promoted after its v^{th} vote. We account for the spread in the number of votes a story has when promoted with a logistic regression to define $P(v)$, in contrast to the prior use of a step function at 40 votes (Hogg and Lerman 2009).

parameter	value
average rate each user visits Digg	$\omega = 0.2\,/\text{hr}$
number of active users	$U = 70,000$
fraction viewing upcoming pages	$c = 0.065$
page view distribution	$\mu = 6.3$, $\lambda = 0.14$
probability a user is a voter's fan	$\rho = 9.48 \times 10^{-6}$
upcoming stories location	$k_{\text{u}} = 3.60\,\text{pages/hr}$
front page location	$k_{\text{f}} = 0.18\,\text{pages/hr}$

Table 1: Model parameters, with times in "Digg hours".

Since we observe votes, not visits to Digg, there is some ambiguity in the values ω and r_F, r_N, which appear only as products in the rate equations. This arbitrary scaling does not affect the relative behavior of fans and non-fans, so for definiteness we pick a specific value for ω. We used the May data to estimate k_{u} and k_{f}, which correspond to 54 and 2.7 stories per hour submitted and promoted, respectively.

We use the non-fan votes for 16 stories in the June data set to estimate c and the "law of surfing" parameters μ and λ by maximum likelihood. We then use fan votes for these stories to evaluate ρ.

Our model involves a population of "active users" during our sample period. We do not observe visits in our data, but can infer the number of active users, U, from the heterogeneity in the number of votes by users. The June data set consists of 16283 users who voted at least once. Fig. 2 shows the distribution of this activity on front page stories. Most users have little activity, suggesting a large fraction of users vote infrequently enough to never have voted during our data sample. Users can be characterized by activity rates. A user with activity rate ν will, on average, vote on νT stories during a sample time T. We model the votes as arising from a Poisson process with mean νT and the heterogeneity arising from a lognormal distribution of user activity rates (Hogg and Szabo 2009). This model gives rise to the extended activity distribution while accounting for the discrete nature of the observations. The latter is important for the majority of users who have low activity rates so will vote only a few times, or not at all, during our sample period. A zero-truncated maximum likelihood estimate (Hilbe 2008) fits this model to the vote distribution of Fig. 2, giving νT lognormally distributed with the mean and standard deviation of $\log(\nu T)$ equal to -2.06 ± 0.03 and 1.82 ± 0.03, respectively. Based on this fit, the curve in Fig. 2 shows the expected number of users with each number of votes. A bootstrap test (Efron 1979) based on the Kolmogorov-Smirnov (KS) statistic shows the vote counts are consistent with this distribution (p-value 0.48). This test and the others reported in this paper account for the fact that we fit the distribution parameters to the data (Clauset, Shalizi, and Newman 2009). This fit indicates about 3/4 of the users had sufficiently low, but nonzero, activity rate that they did not vote during the sample period. We use this value to estimate U.

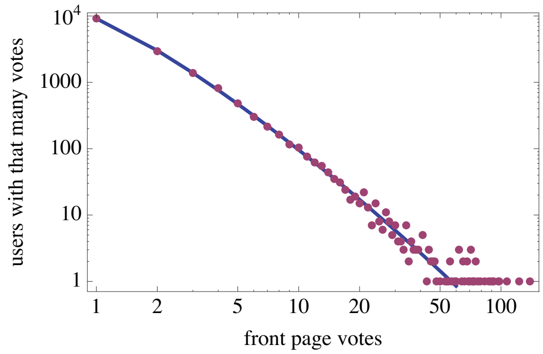

Figure 2: User activity distribution on logarithmic scales. The curve shows the fit to the model described in the text.

Results

We find a wide range of interestingness: the r_N values fit well to a lognormal distribution with mean and standard deviation of $\log(r_N)$ equal to -4.0 ± 0.1 and 0.63 ± 0.07, respectively, with the ranges giving the 95% confidence intervals. A bootstrap test based on the KS statistic shows the r values are consistent with this distribution (p-value 0.1).

The r_F values are approximately lognormally distributed with mean and standard deviation of $\log(r_F)$ equal to -1.8 ± 0.1 and 0.75 ± 0.08, respectively. The KS statistic indicates a weaker fit, with a p-value of 0.04.

We find large variation in the ratio r_F/r_N: ranging from 0 to 87, with median 9.3. The high values correspond to stories that get only a few votes, indicating they are of significantly more interest to the fans of voters than to the general user population, i.e., "niche interest" stories. Overall, there is little relation between how interesting a story is to fans and other users: the correlation between r_F and r_N is -0.11. A randomization test indicates this small correlation is only marginally significant, with p-value 0.05 of arising from uncorrelated values. Stories with high ratios of r_F/r_N tend to be promoted after fewer votes than those stories with low ratios.

Predicting popularity in social media from intrinsic properties of newly submitted content is difficult (Salganik, Dodds, and Watts 2006). However, users' early reactions give some predictability (Hogg and Szabo 2009; Kaltenbrunner, Gomez, and Lopez 2007; Lerman and Galstyan 2008; Szabo and Huberman 2008). As one example, we evaluate whether a story receives at least 500 votes. Table 2 compares the predictions with different methods, including a constrained version of our model with $r_F = r_N$, which assumes no systematic difference in interest between fans and other users. For comparison, direct extrapolation from the v votes observed at early time t gives vt_{final}/t, where we take t_{final} to be 72 hours, a time by which stories have accumulated all, or nearly all, the votes they will ever get. We use a least squares linear fit between these observed and extrapolated values as the prediction. A pairwise bootstrap test indicates the model has a lower prediction error than this extrapolation with p-value of 10^{-2}. This extrapolation method differs from a prior study (Szabo and Huberman 2008) in two ways: 1) we extrapolate from the time required for the story to acquire a given number of votes instead of the number of votes at a given time, and 2) we use early votes after submission (i.e., including when the story is upcoming, where the social network has a large effect) instead of early votes after promotion.

	model		direct
	distinct r	same r	extrapolation
first 216 votes	10%	12%	21%
first 10 votes	18%	23%	29%

Table 2: Prediction errors on whether a story receives at least 500 votes for three methods: 1) the full model which allows distinct values for r_F and r_N, 2) the model constrained to have $r_F = r_N$, and 3) direct extrapolation from the rate the story accumulates votes. This comparison involves 178 promoted stories, of which 137 receive at least 500 votes.

In the case of prediction based on the first 10 votes, which is before the stories are promoted, an additional question is how well the model predicts whether the story will eventually be promoted. We find a 25% error rate in predicting promotion based on the first 10 votes.

We can improve predictions from early votes by using the lognormal distributions of r_F and r_N as the prior probability to combine with the likelihood from the data with Bayes theorem. Using this prior gives little change in r_N, due to the many non-fan votes on each story, but makes large changes in some of the r_F estimates. For example, when predicting based on the first 10 votes, using this prior increases the Spearman rank correlation between predicted and actual number of votes from 0.46 to 0.53. For comparison, this correlation for extrapolation from the first 10 votes is 0.32 and is 0.34 for the model constrained to have $r_F = r_N$. Pairwise bootstrap tests indicate the differences between these values are significant with p-values less than 0.01.

A previous study showed stories initially receiving a small proportion of votes from fans became much more popular than stories which had a high proportion of such votes (Lerman and Galstyan 2008). That work exploited social influence only to make the prediction, and the results were not applicable to stories submitted by poorly connected users which were not quickly discovered by highly connected users. In contrast, the approach described in this paper considers effects of social influence regardless of the connectedness of the submitter, and also accounts for story quality. Fig. 3 shows our model explains this relationship, which arises from the difference in interestingness for fans and non-fans. Specifically, a low fraction of early votes by fans indicates r_N is relatively large to produce the early non-fan votes in spite of the lower visibility of upcoming stories to non-fan users. Once the story is promoted, it then receives relatively more votes from the general user community (most of whom are not fans of prior voters). The separation of effects of visibility and interestingness with our model improves this discrimination compared to just using the raw number of votes by fans and non-fans without regard for the story visibility at the time of the votes. For example, the correlation between the final number of votes

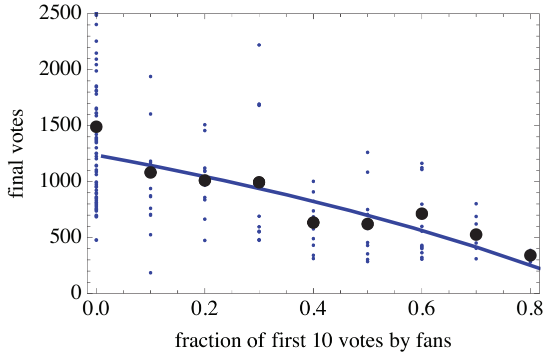

Figure 3: Relation between final number of votes and the fraction of votes by fans among a story's first 10 votes. Small points are individual stories and the large points are the mean values for each number of votes by fans. The curve shows the model prediction.

and r_N/r_F is 0.72 compared to 0.64 for the correlation between the final number of votes and the ratio of non-fan to fan votes.

Discussion

The broad distributions of popularity and user activity on many social media sites can arise from simple macroscopic dynamical rules (Wilkinson 2008). A phenomenological model of the collective attention on Digg describes the distribution of final votes for promoted stories through a decay of interest in news articles (Wu and Huberman 2007). Our models offers an alternative explanation for the vote distribution from the combination of variation in the stories' inherent interest to users and effects of user interface. Crane and Sornette (2008) found dynamics was linked to the quality of videos on YouTube. While these studies aggregated data from tens of thousands of individuals, our method focuses instead on the *microscopic* dynamics, modeling how individual behavior contributes to content popularity.

Early and late popularity are correlated on Slashdot (Kaltenbrunner, Gomez, and Lopez 2007), Digg and YouTube (Szabo and Huberman 2008). The niche interest of content spread mainly via the social network is also seen in Second Life (Bakshy, Karrer, and Adamic 2009).

By accounting for visibility, our model identifies stories of high interest to fans. This could help highlight stories in the friends interface, and recommend new fans to users, based on visibility-adjusted similarity in voting rather than, as commonly done in collaborative filtering (Konstan et al. 1997), just the number of similar votes. We find a wide range of interestingness ratios between fans and non-fans. This explains prior observations that relatively high votes from fans indicate stories are of niche interest. For more precise estimates of interestingness, the web site could track the fraction of users seeing the story that vote for it.

User-contributory web sites typically allow users to designate others whose contributions they find interesting, and the sites highlight the activity of linked users. Thus our stochastic model, explicitly distinguishing behavior of users based on whether they are linked to users who submitted or previously rated the content, could apply to many such web sites.

References

Bakshy, E.; Karrer, B.; and Adamic, L. A. 2009. Social influence and the diffusion of user-created content. In *Proc. of the 10th ACM Conf. on Electronic Commerce (EC09)*, 325–334. NY: ACM.

Clauset, A.; Shalizi, C. R.; and Newman, M. E. J. 2009. Power-law distributions in empirical data. *SIAM Review* 51:661–703.

Crane, R., and Sornette, D. 2008. Viral, quality, and junk videos on YouTube: Separating content from noise in an information-rich environment. In Lerman, K., et al., eds., *Proc. of the AAAI Symposium on Social Information Processing*, 18–20.

Efron, B. 1979. Bootstrap methods: Another look at the jackknife. *Annals of Statistics* 7:1–26.

Hilbe, J. M. 2008. *Negative Binomial Regression*. Cambridge Univ. Press.

Hogg, T., and Lerman, K. 2009. Stochastic models of user-contributory web sites. In *Proc. of the Third International Conference on Weblogs and Social Media (ICWSM2009)*, 50–57. AAAI.

Hogg, T., and Szabo, G. 2009. Diversity of user activity and content quality in online communities. In *Proc. of the Third International Conference on Weblogs and Social Media (ICWSM2009)*, 58–65. AAAI.

Huberman, B. A.; Pirolli, P. L. T.; Pitkow, J. E.; and Lukose, R. M. 1998. Strong regularities in World Wide Web surfing. *Science* 280:95–97.

Kaltenbrunner, A.; Gomez, V.; and Lopez, V. 2007. Description and prediction of slashdot activity. In *Proc. 5th Latin American Web Congress (LA-WEB 2007)*.

Konstan, J. A.; Miller, B. N.; Maltz, D.; Herlocker, J. L.; Gordon, L. R.; and Riedl, J. 1997. GroupLens: Applying collaborative filtering to Usenet news. *Communications of the ACM* 40(3):77–87.

Lerman, K., and Galstyan, A. 2008. Analysis of social voting patterns on Digg. In *Proceedings of the 1st ACM SIGCOMM Workshop on Online Social Networks*.

Lerman, K. 2007. Social information processing in social news aggregation. *IEEE Internet Computing: special issue on Social Search* 11(6):16–28.

Salganik, M. J.; Dodds, P. S.; and Watts, D. J. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *Science* 311:854–856.

Szabo, G., and Huberman, B. A. 2008. Predicting the popularity of online content. Technical report, HP Labs. Available at hpl.hp.com/research/scl/papers/predictions.

Wilkinson, D. M. 2008. Strong regularities in online peer production. In *Proc. of the 2008 ACM Conference on E-Commerce*, 302–309.

Wu, F., and Huberman, B. A. 2007. Novelty and collective attention. *Proc. of the Natl. Acad. Sci.* 104:17599–17601.