

Social Dynamics of Digg

Tad Hogg · Kristina Lerman

the date of receipt and acceptance should be inserted later

Abstract Online social media provide multiple ways for people to find interesting content submitted by other users. One important method is highlighting content that is highly rated by similar users, whose similarity is indicated implicitly by their behavior or explicitly by links in a social network. By selectively presenting content to users, and aggregating over their preferences, social media sites attempt to identify generally interesting content. Using data from one such site, the news aggregator Digg, we use a stochastic model of user behavior to distinguish the effect of the increased visibility from the network from how interested users are in the content. We find a wide range of interest, identifying stories primarily of interest to users in the network from those of more general interest to the user community. We show how to use this model to predict a story's eventual popularity from users' early reactions to the story, and estimate the reliability of the prediction for individual stories.

Keywords social media · modeling · stochastic modeling · information diffusion

T. Hogg
Institute for Molecular Manufacturing
Palo Alto, CA 94301
E-mail: tadhogg@yahoo.com

K. Lerman
USC Information Sciences Institute
Marina del Rey, CA 90292, USA
Tel.: 310-448-8714
Fax: 310-822-0751
E-mail: lerman@isi.edu

1 Introduction

The explosive growth of the Social Web hints at collective problem-solving made possible when people have tools to connect, create and organize information on a massive scale. The social news aggregator Digg, for example, allows people to collectively identify interesting news stories. The microblogging service Twitter has created a cottage industry of third-party applications, such as identifying trends from the millions of conversations taking place on the site and notifying you when your friends are nearby. Other sites allow like-minded people to collectively create encyclopedias, develop software, and invest in social causes. Analyzing records of complex social activity can identify communities and important individuals within them [4], suggest relevant readings [21], and identify events [23] and trends [18,26].

Effective use of this technology requires understanding how the social dynamics emerges from the decisions made by interconnected individuals. Stochastic modeling framework offers one tool to study this problem. Such models represent each user as a stochastic process with a few states, e.g., a simple Markov processes whose future state depends only on its present state and the input it receives. Ref. [17,11] used this approach to study voting on the social news aggregator Digg and showed qualitative agreement between the model and voting patterns of Digg’s users. However, quantitative evaluation of the model was limited by the poor quality of data, which was extracted by scraping Digg’s web pages and contained extraction mistakes and missing data. These introduced errors into model parameter estimation process, thereby limiting the range of behaviors the model could identify. We, on the other hand, study a complete voting record of a subset of stories retrieved using Digg API. This high quality data enabled us to discover and measure systematic differences among classes of users and iteratively refine the model to include new aspects of the Digg user interface.

In this paper we propose two refinements to the existing modeling approach. First, we explicitly allow for systematic differences in *interest* in news stories for linked and unlinked users. This enables us to model a key aspect of social media: the extent to which links indicate commonality of interests of users. We also introduce additional aspects of the Digg user interface in the model, to account for cases where parameter estimation identified anomalous behaviors in the existing model. As the second major contribution, we describe the methodology to measure confidence intervals, which allow us to evaluate the quality of the model’s prediction of user behavior. We show that confidence intervals are highly correlated with the error between the predicted and actual votes stories accrue.

This paper is organized as follows. The next section describes Digg and our data set. We then present a stochastic model of user behavior on Digg that explicitly incorporates different behaviors depending on social network links created by the users. Using this model, we quantify the nature of these differ-

ences and discuss how the model can help predict eventual popularity of newly submitted content from early reaction by a few users and their relationships in the social network. Finally, we compare our approach with other studies and discuss possible applications of stochastic models incorporating social network structure.

2 Digg: A Social News Portal

With over 3 million registered users, the social news aggregator Digg is a popular news portal. Digg allows users to submit and rate news stories by voting on, or ‘digging’, them. Every day Digg promotes a few percent of submitted stories to the highly visible *front page*. Although the exact promotion mechanism is kept secret and changes occasionally, it appears to take into account the number of votes the story receives. Digg’s popularity is fueled in large part by the emergent front page, which is created by the collective decisions of its many users.

2.1 User interface

Submitted stories appear in the *upcoming* stories list, where they remain for about 24 hours or until promoted to the front page. By default, Digg shows upcoming and front page stories in recency lists i.e., in reverse chronological order with the most recently submitted (promoted) story at the top of the list. However, a user may choose to display stories by popularity or by some broad topic. Popularity lists show stories with the most votes up to that time, e.g., the most popular stories submitted (promoted) in the past day or week. Each list is divided into pages, with 15 stories on each page, and the user has to click to see subsequent pages.

Digg allows users to designate friends and track their activities. The friend relationship is asymmetric. When user A lists user B as a *friend*, A can follow the activities of B but not vice versa. We call A the *fan*, or follower, of B . Specifically, Digg’s friends interface shows users the stories their friends recently submitted or voted for.¹

In this paper, we focus on the recency and “popular in the last 24 hours” lists for all stories and the friends interface list for each user. These lists appear to account for most of the votes a story receives.

¹ At the time of data collection Digg offered a social filtering feature which recommended stories, including upcoming stories, that were liked by users with a similar voting history. It is not clear how frequently users employed these features and we do not explicitly include them in our model.

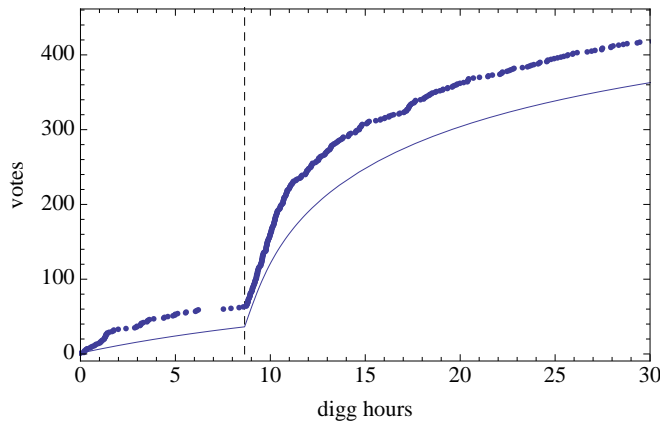


Fig. 1 Voting behavior: the number of votes vs. time, measured in Digg hours, for a promoted story. The curve shows the corresponding solution from our model and the dashed vertical line indicates when the story was promoted to the front page. This story eventually received 452 votes.

2.2 Evolution of story popularity

Most Digg users focus on front page stories, so upcoming stories accrue votes slowly. When a story is promoted, it becomes visible to many more users and accrues votes rapidly. Fig. 1 shows the evolution of the number of votes for a story submitted in June 2009. The point where the slope abruptly increases corresponds to promotion to the front page. As the story ages, accumulation of new votes slows down, and after a few days stories typically no longer receive additional 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. Stories that are never promoted receive few votes, in many cases just a single vote from the submitter, and are removed after about 24 hours.

A challenge for understanding this variation in popularity is the interaction between the stories' *visibility* (how Digg displays them) and their *interestingness* to users. Models accounting for the structure of the Digg interface can help separate and evaluate these contributions to story popularity.

2.3 Data set

We used Digg API to collect complete (as of July 2, 2009) voting histories of all stories promoted to the front page of Digg in June 2009. For each story, we collected story id, submitter's id, list of voters with time of each vote. We also collected the time each story was promoted to the front page. In total, the data set contains over 3 million votes on 3,553 promoted stories. We did not retrieve

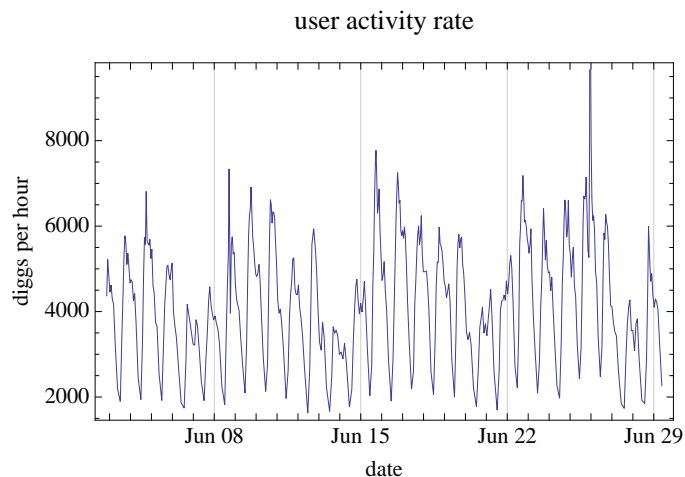


Fig. 2 Voting rate (digs per hour) on front page stories during June 2009. The indicated dates are the start of each day (0:00 GMT). The minimum in daily activity is around noon GMT.

data about stories that were submitted to Digg during that time period but were never promoted. Thus our focus with this data is on the behavior of promoted stories, which receive most of the attention from Digg users.

We define an *active user* as any user who voted for at least one story on Digg during the data collection period. Of the 139,409 active users, 71,367 designated at least one other user as a friend. We extracted the friends of these users and reconstructed the fan network of active users, i.e., a directed graph of active users who are following activities of other users.

Over the period of a month, some of the voters in our sample deleted their accounts and were marked “inactive” by Digg. Such cases represent a tiny fraction of all users in the data set; therefore, we take the number of users to be constant.

2.4 Daily activity variation

Activity on Digg varies considerably over the course of a day, as seen in Fig. 2. Adjusting times by the cumulative activity on the site accounts for this variation and improves predictions. Following [26,12] we define the “Digg time” between two events (e.g., votes on a story) as the total number of votes made during the time between those events. Note, that we have collected only votes on stories that were eventually promoted to the front page. In our data set, there are on average about 4000 votes on front page stories per hour, with a range of about a factor of 3 in this rate during the course of a day. This behavior is similar to that seen in an extensive study of front page activity in

2007 [26], and as in that study we scale the measure by defining a “Digg hour” to be the average number of front page votes in an hour.

3 Social Dynamics of Digg

A key issue in designing stochastic models is finding a useful combination of simplicity, accuracy and available data to calibrate the model, i.e., determine its parameters. Stochastic models of online social media describe the joint behavior of users and content on the web site. Since these sites receive much more contributed content than users have time or interest to examine, one important property to model is how readily users can find content. A second key property is how users react to content once they find it. Thus an important modeling choice for social media is the level of detail sufficient to distinguish user behavior and content visibility. Following the practice of population dynamics [8] and epidemic modeling [9] we group users and content into compartments, and assume that individuals within each group or compartment behave in a sufficiently similar manner that their differences do not affect the main questions of interest in developing the model. In the case of Digg, one such question is the number of votes a story receives over time. In our approach, we treat stories independently and focus on how a single story accumulates votes, based on the combination of how easily users can find the story and how interesting it is to different groups of users. If we find that the coarse groupings do not explain some observed behavior, we refine the groups as needed to account for this behavior.

Following Ref. [17, 11], we start with a simple model in which story visibility is determined primarily by its location on the recency and friends lists, and use a single value to describe the story’s interestingness to the user community. As users submit new stories or Digg promotes stories to the front page, the location of the given story on the recency list changes in a predictable way. We use the “law of surfing” [14] to relate location of the story to how readily users find it. This model successfully captured the qualitative behavior of typical stories on Digg and how that behavior depended on the number of fans of the story’s submitter [11, 19].

However, the simple model did not quantitatively account for a number of behaviors identified in the new data set. These included the significant daily variation in activity rates seen in Fig. 2 and systematic differences in behavior between fans of a story’s submitter and other users. In particular, the new data was sufficiently detailed to show users tend to find stories their friends submit as more interesting than stories friends vote on but did not submit. A second issue with the earlier model arose from the unexpectedly large number of votes accrued by stories when they were far down the recency list, especially for upcoming stories where the large rate of new submissions means a given story remains near the top of the recency list for only a few minutes. In order to account for such votes, the model estimated “the law of surfing” parameters

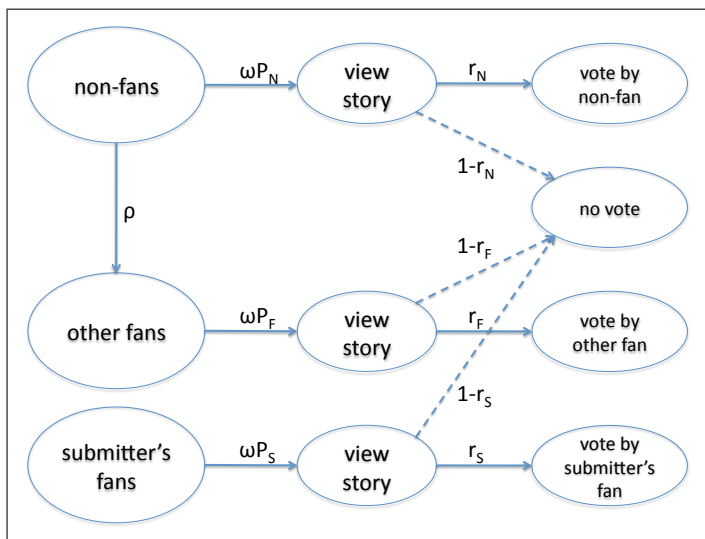


Fig. 3 State diagram for a user.

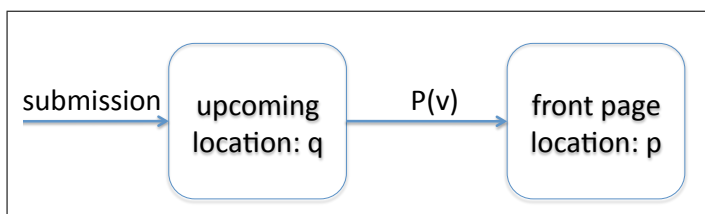


Fig. 4 State diagram for a story. A story not promoted after sufficient time (usually within a day) is removed (a state transition not shown in the diagram).

which would have users browse an implausibly large number of pages while visiting Digg.

These observations suggested the more elaborate model described in this paper, where we consider systematic differences in interestingness between fans and other users and include additional ways Digg makes stories visible to users.

3.1 User model

We allow for differences between users by separating them into groups, and assume that visibility of stories and voting behavior of users within each group is statistically the same. That is, users within a group share the same parameter values in the model. We refine the previous stochastic model of Digg [11] by not only distinguishing votes from fans and non-fans [12], but also allowing for differences between fans of the *submitter* and fans of *other* voters who are

not also fans of the submitter. The state diagram Fig. 3 shows the compartmentalized user model. The state **submitter's fans** includes all users who are fans of the submitter and have not yet seen the story; the **other fans** state includes all users who are fans of other voters but not the submitter, who have not yet seen the story; and the **non-fans** state includes all users who are neither fans of the submitter nor other voters, and have not seen the story yet. The state **no vote** includes all users who have seen the story and decided not to vote for it. With respect to votes on a given story treated in this model, users visit Digg according to a Poisson process with average rate ω in terms of Digg time.

Users transition between states stochastically by browsing Digg's web pages and voting for stories. The submitter provides a story's first vote. All of her fans start in the **submitter's fans** state, and all other users start in the **non-fans** state. Each vote causes non-fan users who are that voter's fans and who have not yet seen the story to transition from the **non-fans** state to the **other fans** state. A user making this transition is not aware of that change until later visiting Digg and seeing the story on her friends list, i.e., this transition is due to actions of other users.

Once a user sees a story, she will vote for it with probability given by how interesting she finds the story. Nominally people become fans of those whose contributions they consider interesting, suggesting fans likely have a systematically higher interest in stories than non-fans. Our model accounts for this possibility by having three different *story interestingness* parameters: r_S , r_F and r_N set the probability a user who is a fan of the submitter, previous voters, or a non-fan respectively will vote for the story given she sees it. Users in each category also have a different probability to see stories, which is determined by the story's *visibility* to that category of users. Users vote at most once on a story, and our focus is on the final decision to vote or not after the user sees the story.

The visibility of stories changes in time as stories age and accrue votes. The state diagram of stories is shown in Fig. 4. A story starts at the top of the upcoming pages, with location $q = 1$. The location increases with each new submission. Digg selects a small fraction of submitted stories to promote to the front page. Stories not promoted after a day or so are removed. A promoted story starts at the top of the front pages, with location $p = 1$. The location increases as additional stories are promoted.

These state diagrams lead to a description of the average rates of growth [20] for votes from submitter fans, other fans and non-fans of prior voters, v_S , v_F and v_N , respectively:

$$\frac{dv_S}{dt} = \omega r_S P_S S \quad (1)$$

$$\frac{dv_F}{dt} = \omega r_F P_F F \quad (2)$$

$$\frac{dv_N}{dt} = \omega r_N P_N N \quad (3)$$

where t is the Digg time since the story's submission and ω is the average rate a user visits Digg (measured as a rate per unit Digg time). v_N includes the story's submitter. P_S , P_F and P_N denote the story's *visibility* and r_S , r_F and r_N denote the story's *interestingness* to users who are submitter fans, other fans or non-fans 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.

These voting rates depend on the number of users in each category who have not yet seen the story: S , F and N . The quantities change as users see and vote on the story according to

$$\frac{dS}{dt} = -\omega P_S S \quad (4)$$

$$\frac{dF}{dt} = -\omega P_F F + \rho N \frac{dv}{dt} \quad (5)$$

$$\frac{dN}{dt} = -\omega P_N N - \rho N \frac{dv}{dt} \quad (6)$$

with $v = v_S + v_F + v_N$ the total number of votes the story has received. The quantity 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. For simplicity, we treat this probability as a constant over the voters, thus averaging over the variation due to clustering in the social network and the number of fans a user has. The first term in each of these equations is the rate the users see the story. The second terms arise from the rate the story becomes visible in the friends interface of users who are not fans of previous voters but are fans of the most recent voter.

Initially, the story has one vote (from the submitter) and the submitter has S_0 fans, so $v_S(0) = v_F(0) = 0$, $v_N(0) = 1$, $S(0) = S_0$, $F(0) = 0$ and $N(0) = U - S_0 - 1$ where U is the total number of active users at the time the story is submitted. Over time, a story becomes less visible to users as it moves down the upcoming or (if promoted) front page lists, thereby attracting fewer votes and hence fewer new fans of prior voters. If the story gathers many votes, it moves to the front of the popularity list, so becomes more visible.

We find only a small correlation between voting activity and the number of fans. Thus we consider the average rate all users visit Digg, denoted by ω , rather than having the rate depend on the number of fans a user has.

Fig. 5 shows the range of votes the stories receive by 24 hours after promotion. Generally, stories have most votes from non-fans, somewhat fewer from other fans and a relatively small number from submitter's fans. The number of votes from submitter's fans is weakly correlated with the numbers from other fans (correlation coefficient 0.09) and non-fans (0.05). The numbers from other fans and non-fans are highly correlated (0.90).

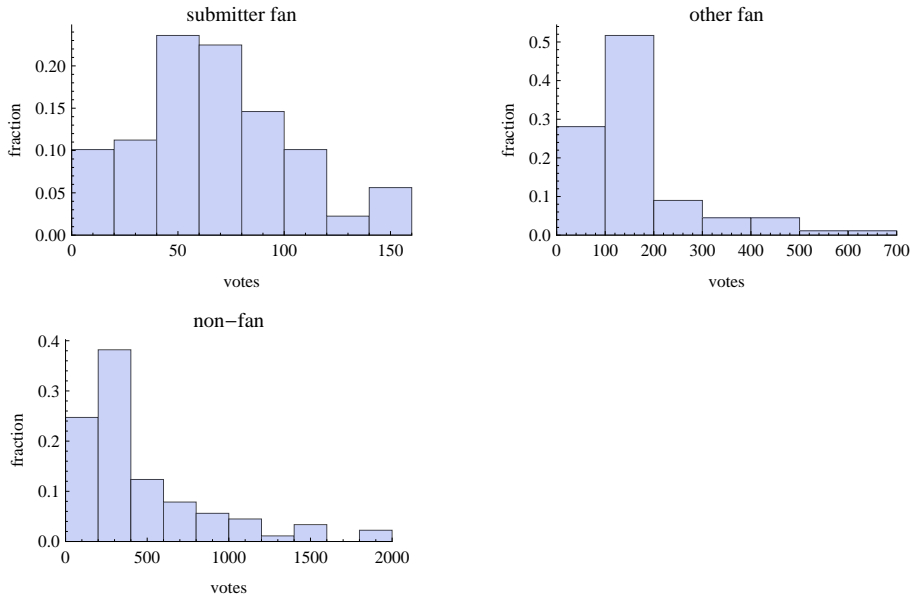


Fig. 5 Distribution of each type of vote a story receives 24 hours after promotion.

3.2 Story visibility

A fan easily sees the story via the friends interface, so we take $P_S = P_F = 1$ for front page stories. While the story is upcoming, it appears in the friends interface but users do not necessarily choose to see upcoming stories friends liked. Users can readily make this distinction because the friends interface distinguishes upcoming from front page stories. We characterize the lower visibility of upcoming stories with constants c_S and c_F which are less than 1. The corresponding visibility is then $P_S = c_S$ and $P_F = c_F$.

Users who are not fans of prior voters must find the story on the front or upcoming pages. Thus P_N depends on how users navigate through these pages and the story’s location at the time the user visits Digg. This navigation is not given by our data. Instead, we use a model of how users navigate through a series of web pages. Typically, successively smaller fractions of users visit later pages in lists presented on a sequence of web pages. One model of this behavior considers users estimating the value of continuing at the site, and leaving when that value becomes negative [14]. This “law of surfing” leads to an inverse Gaussian distribution of the number of pages m a user visits before leaving the web site,

$$e^{-\frac{\lambda(m-\mu)^2}{2m\mu^2}} \sqrt{\frac{\lambda}{2\pi m^3}} \quad (7)$$

with mean μ and variance μ^3/λ . We use this distribution for user navigation on Digg [11].

The visibility of a story on the m^{th} front or upcoming page is the fraction of users who visit *at least* m pages, i.e., the upper cumulative distribution of Eq. (7). For $m > 1$, this fraction is

$$f_{\text{page}}(m) = \frac{1}{2} \left(F_m(-\mu) - e^{2\lambda/\mu} F_m(\mu) \right) \quad (8)$$

where $F_m(x) = \text{erfc}(\alpha_m(m-1+x)/\mu)$, erfc is the complementary error function, and $\alpha_m = \sqrt{\lambda/(2(m-1))}$. For $m = 1$, $f_{\text{page}}(1) = 1$. The visibility of stories decreases in two distinct ways when a new story arrives. First, a story moves down the list on its current page. Second, a story at the 15th position moves to the top of the next page. For simplicity, we model these processes as decreasing visibility in the same way through m taking on fractional values within a page, e.g., $m = 1.5$ denotes the position of a story half way down the list on the first page.

Digg presents several lists of stories. We focus on two lists as the major determinants of visibility for front page stories: reverse chronological order (“recency”) and most popular in the past 24 hours (“popularity”). Users can also find stories via other means. For instance, Digg includes other lists showing recent and popular stories in specific topics (e.g., sports or business) and popularity over longer time periods, e.g., the previous week. Stories on Digg may also be linked to from external web sites (e.g., the submitter’s blog).

For front page votes, the recency and popularity lists provide the bulk of non-fan votes while the stories are close to the top of at least one of these lists, as illustrated in Fig. 6. Here rank on the recency list at the time of a vote is the number of stories promoted more recently than that story and the location on the popularity list is the number of stories, promoted within the 24 hours prior to the vote, with more votes. In each case, a page shows 15 stories, so the location in terms of number of pages, as shown in the figure, is $1/15^{\text{th}}$ the rank, starting from page 1. Some votes occur while stories are far down both the recency and popularity lists, so the user likely finds the story by another method.

Thus, for users who are not fans of prior voters, we take into account three ways for them to find a story on Digg: via the recency list, via the popularity list or via one of the other methods described above. We combine visibility from these three methods assuming independent choices by users, giving the probability to see the story as

$$P_{\text{visibility}}(t, v) = 1 - (1 - f_{\text{page}}(p(t))) (1 - f_{\text{page}}(p_{\text{popularity}}(v))) (1 - \beta) \quad (9)$$

where $p(t)$ and $p_{\text{popularity}}(v)$ are the locations of the story on the recency and popularity lists, respectively, and β is the probability to find the story by another method. Although the positions of the stories on these lists depend on the specific stories submitted or promoted shortly after the story, these locations are approximately determined by the time t and number of votes v the story has, as described below. For visibility by other methods, we simply

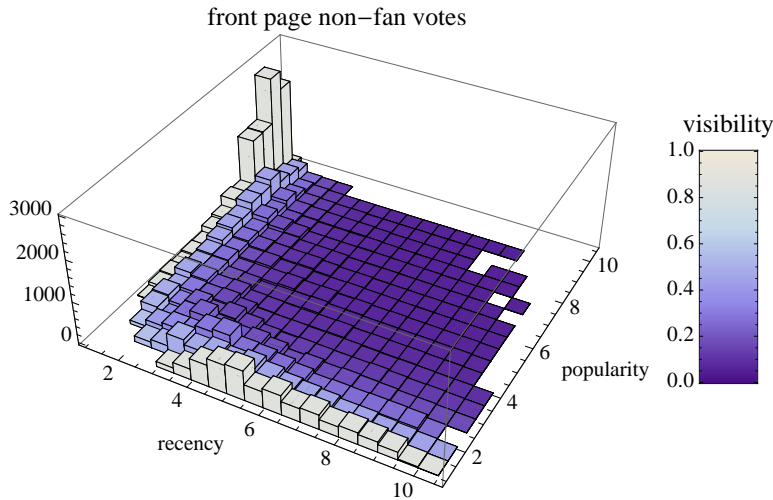


Fig. 6 Distribution of front page non-fan votes by location of the story on recency and popularity lists (for votes within 24 hours of story promotion), for a sample of 100 stories with a total of 41615 such votes. The colors indicate the visibility for each location predicted by the model parameters using Eq. (9), ranging between 0 and 1 as indicated on the legend.

take β to be a constant, independent of story properties such as time since submission or number of votes. That is, we do not explicitly model factors affecting the visibility of stories by other methods, as the recency and popularity lists account for the bulk of the non-fan votes determined by our parameter estimates discussed below.

In general, the location of a story on the recency and popularity lists could be viewed as additional state variables for the story, which change as new stories are added and gain votes. Instead of modeling this in detail, we find a close relation between location and time (for recency) or votes (for popularity). Thus instead of additional state variables, we approximate these locations based on time and votes.

Position of a story in the recency list Using Digg time to account for the daily variation in activity on the site, the rate of story submission and promotion is close to linear. This behavior gives a simple relation between story location on the recency list and time. Specifically, the page number of a story on the upcoming or front page lists is [11]

$$p(t) = \begin{cases} k_u t + 1 & \text{if } t < T_{\text{promotion}} \\ k_f (t - T_{\text{promotion}}) + 1 & \text{otherwise} \end{cases} \quad (10)$$

where $T_{\text{promotion}}$ is the time the story is promoted to the front page and the slopes are given in Table 1. Since each page holds 15 stories, these rates are $1/15^{\text{th}}$ the story submission and promotion rates, respectively. In our data set, there are about 21,000 stories submitted per day, and 110 promoted per day.

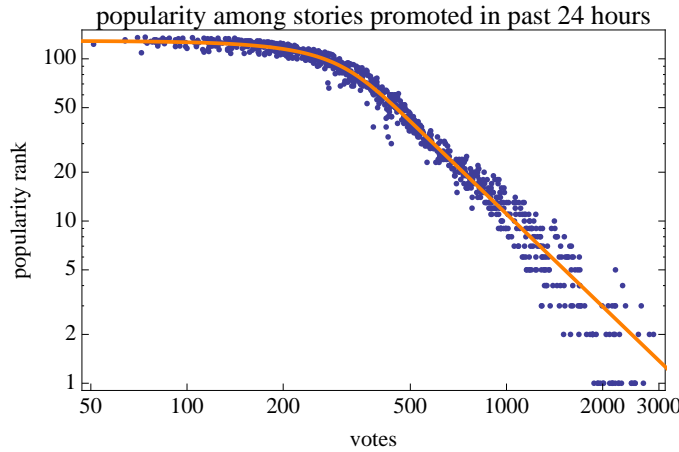


Fig. 7 Relation between rank on popularity list and number of votes for front page stories on a log-log plot. The curve is the fit to a double-Pareto lognormal distribution.

Position of a story in the popularity list The position of a story on the popularity list is the number of stories submitted or promoted in the previous 24 hours with more votes, for stories on upcoming or front page lists, respectively. The distribution of votes among stories in a 24 hour period is similar from day to day. Thus a story's position on the popularity list, determined by the location of its number of votes in this distribution, is approximately a function of its number of votes alone, with only minor variation depending on the time (i.e., the set of other stories from the past 24 hours). Thus we model the position as depending only on the number of votes the story has at the time, i.e., consider $p_{\text{popularity}}(v)$ as a function of the number of votes the story has. This gives a simple, approximate relation between the actual location and number of votes – ignoring the minor variations due to the specific stories promoted at different times.

Fig. 7 shows the relation between popularity rank and number of votes for a sample of front page votes within 24-hours of story promotion. To identify a suitable functional approximation for this relation, we note that a story typically accumulates votes at a rate proportional to how interesting it is to the user population. As seen in prior analysis of votes in 2006 on Digg [11], we expect the interestingness to be distributed according to a lognormal distribution. Thus if we observe a sample of votes on stories over a the same time interval for each story, the distribution of votes, and hence location on the popularity list, would follow a lognormal distribution. However, the popularity list includes stories of various times up to 24 hours since submission or promotion. Thus some stories of high interest will have few votes because they were just recently submitted or promoted, and conversely some stories with only moderate interestingness will have relatively many votes because they have been available for votes for nearly 24 hours. The combination of lognormal distribution of rates for accumulating votes and this variation in

the observation times modifies the tails of the lognormal to be power-law, i.e., a double-Pareto lognormal distribution [24].

We find such a distribution provides a good fit to the observed positions on the front page popularity list, as indicated in Fig. 7. The fit for the rank (number of stories above the given one in the popularity list, so story promoted in the past 24 hours with the most votes has rank 0) is

$$\text{rank} = S(1 - \Lambda(a, b, \nu, \sigma; v)) \quad (11)$$

where $S = 129.0 \pm 0.1$ is the average number of stories promoted in 24 hours and $\Lambda(\dots; v)$ is the cumulative distribution of a double-Pareto lognormal distribution, i.e., fraction of cases with fewer than v votes. The parameters $a = 1.90 \pm 0.005$ and $b = 2.50 \pm 0.03$ are the power-law exponents for the upper and lower tails of the distribution, respectively, and the parameters $\nu = 5.88 \pm 0.002$ and $\sigma = 0.16 \pm 0.004$ characterize the location and spread of the lognormal behavior in the center of the distribution. In particular, this fit captures the power-law tail relating stories near the top of the popularity list with the number of votes the story has. These are the cases for which the popularity list contributes significantly to the overall visibility of a story. More precisely, the Kolmogorov-Smirnov (KS) statistic shows the vote counts are consistent with this distribution (p -value 0.92). We use this distribution, combined with the rate stories are promoted, to relate the number of votes a story has to its position on the popularity list, providing a functional form for $p_{\text{popularity}}(v)$.

The popularity rank for upcoming stories submitted in the past 24 hours is more complicated than for the front pages due to the promotion. Stories with many votes are more likely to be promoted, and hence removed from the popularity list for upcoming stories. This removal alters the upper tail of the distribution and hence the numbers of votes for stories appearing near the top of the popularity list. Moreover, out of the ≈ 21000 stories submitted each day, our data includes only the ≈ 100 stories per day eventually promoted. However, popularity significantly contributes to visibility only for stories near the top of the popularity list. Thus for our model, it is sufficient to determine the relation between votes and rank for upcoming stories with relatively many votes. Such stories are likely to be promoted eventually and hence included in our sample. Instead of a power-law tail, our data on the eventually promoted stories is better fit by an exponential for the upcoming stories with relatively many votes, and hence near the top of the popularity list:

$$\text{rank} = e^{c-dv} \quad (12)$$

with $c = 5.3 \pm 0.01$ and $d = 0.029 \pm 0.0002$. This fits well for upcoming stories submitted within the past 24 hours with more than 100 votes, corresponding to rank of about 20 or less on the upcoming popularity list. For stories with few votes, e.g., fewer than 10 or 20, this fit based on the ≈ 100 stories per day eventually promoted substantially underestimates the rank. Nevertheless,

the estimated rank for such stories is still sufficiently large that the law of surfing parameters we estimate indicate users do not find such stories via the popularity list. Thus this underestimate does not significantly affect our model's behavior for upcoming stories.

Friends interface The fans of the story's submitter can find the story via the friends interface. As additional people vote on the story, their fans can also see the story. We model this with $F(t)$, the number of fans of voters on the story by time t who are not also fans of the submitter and have not yet seen the story. Although the number of fans is highly variable, we use the average number of additional fans from an extra vote, ρN , in Eq. (4).

4 Parameter estimation

We estimate model parameters using 100 stories selected from the middle of our sample, which cover about one day of promoted stories.

4.1 Estimating parameters from observed votes

In our model, story location affects visibility only for non-fan voters since fans of prior voters see the story via the friends interface. Thus we use just the non-fan votes to estimate visibility parameters, via maximum likelihood. Specifically, we use the non-fan votes to estimate the "law of surfing" parameters μ and λ , as well as the probability for finding the story some other way, β . Separating votes by the different interfaces by which users find stories provides more precise estimation than the prior model [11].

This estimation involves comparing the observed votes to the voting rate from the model. As described above, the model uses rate equations to determine the average behavior of the number of votes. A simple approach to relate this average to the observed number of votes is to assume the votes from non-fan users form a Poisson process whose expected value is $dv_N(t)/dt$, given by Eq. (3). This rate changes with time and depends on the model parameters.

For a Poisson process with a constant rate v , the probability to observe n events in time T is the Poisson distribution $e^{-vT}(vT)^n/n!$. This probability depends only on the *number* of events, not the specific times at which they occur. Estimating a constant rate involves maximizing this expression, giving $v = n/T$. Thus the maximum-likelihood estimate of the rate for a constant Poisson process is equal to the average rate of the observed events.

In our case, the voting rate changes with time, requiring a generalization of this estimation. Specifically consider a Poisson process with nonnegative rate $v(t)$ which depends on one or more parameters to be estimated. Thus in

a small time interval $(t, t + \Delta t)$, the probability for a vote is $v(t)\Delta t$, and this is independent of votes in other time intervals, by the definition of a Poisson process. Suppose we observe n votes at times $0 < t_1 < t_2, \dots < t_n < T$ during an observation time interval $(0, T)$. Considering small time intervals Δt around each observation, the probability of this observation is

$$\begin{aligned} & P(\text{no vote in } (0, t_1))v(t_1)\Delta t \times \\ & P(\text{no vote in } (t_1, t_2))v(t_2)\Delta t \times \\ & \quad \dots \\ & P(\text{no vote in } (t_{n-1}, t_n))v(t_n)\Delta t \times \\ & P(\text{no vote in } (t_n, T)) \end{aligned}$$

The probability for no vote in the interval (a, b) is

$$\exp\left(-\int_a^b v(t)dt\right)$$

Thus the log-likelihood for the observed sequence of votes is

$$-\int_0^T v(t)dt + \sum_i \log v(t_i) \quad (13)$$

The maximum-likelihood estimation for parameters determining the rate $v(t)$ is a trade-off between these two terms: attempting to minimize $v(t)$ over the range $(0, T)$ to increase the first term while maximizing the values $v(t_i)$ at the specific times of the observed votes. If $v(t)$ is constant, this likelihood expression simplifies to $-vT + n \log v$ with maximum at $v = n/T$ as discussed above for the constant Poisson process. When $v(t)$ varies with time, the maximization selects parameters giving relatively larger $v(t)$ values where the observed votes are clustered in time.

We combine this log-likelihood expression from the votes on several stories, and maximize the combined expression with respect to the story-independent parameters of the model, with the interestingness parameters determined separately for each story.

4.2 User activity

Our model involves a population of active users who visit Digg during our sample period and vote on stories. Specifically, the model uses the rate users visit Digg, ωU . We do not observe visits in our data, but can infer the relevant number of active users, U , from the heterogeneity in the number of votes by users. The data set consists of 139409 users who voted at least once during the sample period, giving a total of 3018197 votes. Fig. 8 shows the distribution

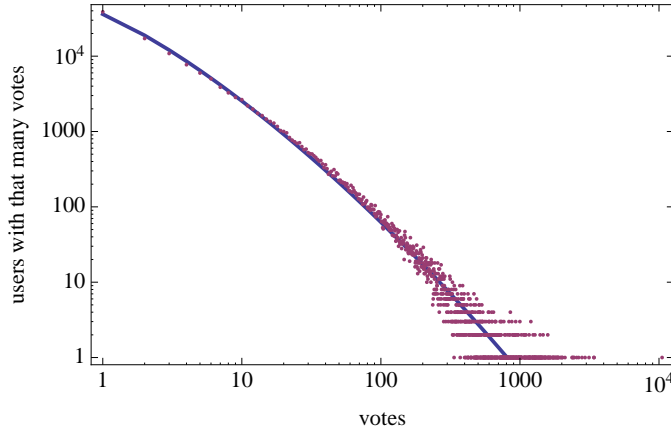


Fig. 8 User activity distribution on logarithmic scales. The curve shows the fit to the model described in the text.

of this activity. Most users have little activity during the sample period, suggesting a large fraction of users vote infrequently enough to never have voted during the time of our data sample. This behavior can be characterized by an activity rate for each user. A user with activity rate ν will, on average, vote on νT stories during a sample time T . We model the observed votes as arising from a Poisson process whose expected value is νT and the heterogeneity arising from a lognormal distribution of user activity rates [13]:

$$P_{\text{lognormal}}(\mu, \sigma; r) = \frac{1}{\sqrt{2\pi} r \sigma} \exp\left(-\frac{(\mu - \log(r))^2}{2\sigma^2}\right) \quad (14)$$

where parameters μ and σ are the mean and standard deviation of $\log(r)$.

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.

Specifically, for n_k users with k votes during the sample period, this mixture of lognormal and Poisson distributions [3, ?] gives the log-likelihood of the observations as

$$\sum_k n_k \log P(\mu, \sigma; k)$$

where $P(\mu, \sigma; k)$ is the probability of a Poisson distribution to give k votes when its mean is chosen from a lognormal distribution $P_{\text{lognormal}}$ with parameters μ and σ . From Eq. (14),

$$P(\mu, \sigma; k) = \frac{1}{\sqrt{2\pi}\sigma k!} \int_0^\infty \rho^{k-1} e^{-\frac{(\log(\rho)-\mu)^2}{2\sigma^2} - \rho} d\rho$$

for integer $k \geq 0$. We evaluate this integral numerically. In terms of our model parameters, the value of μ in this distribution equals νT .

Since we don't observe the number of users who did not vote during our sample period, i.e., the value of n_0 , we cannot maximize this log-likelihood expression directly. Instead, we use a zero-truncated maximum likelihood estimate [10] to determine the parameters μ and σ for the vote distribution of Fig. 8. Specifically, the fit is to the probability of observing k votes conditioned on observing at least one vote. This conditional distribution is $P(\mu, \sigma; k)/(1 - P(\mu, \sigma; 0))$ for $k > 0$, and the corresponding log-likelihood is

$$\sum_{k>0} n_k \log P(\mu, \sigma; k) - U_+ \log(1 - P(\mu, \sigma; 0))$$

where U_+ is the number of users with at least one vote in our sample. Maximizing this expression with respect to the distribution's parameters μ and σ gives νT lognormally distributed with the mean and standard deviation of $\log(\nu T)$ equal to -0.10 ± 0.04 and 2.43 ± 0.02 , respectively. Based on this fit, the curve in Fig. 8 shows the expected number of users with each number of votes. This is a discrete distribution: the lines in the figure between the expected values serve only to distinguish the model fit from the points showing the observed values.

With these estimated parameters, $P(\mu, \sigma; 0) = 0.43$, indicating 43% 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 , the number of active users during our sample period: $U = U_+/(1 - P(\mu, \sigma; 0))$.

4.3 Links among users

We observe $u = 258,218$ users with fans, and these users have a total of $c = 1,731,658$ connections. Our data has 139,409 distinct voters, of which 78,007 have no fans. We observe little correlation between links and voting activity, so estimate the fraction of users with zero fans from the ratio of these values, i.e., about 56%. Thus the average number of fans per user, including users without fans, is $c/(1.56u) \approx 4.3$.

Our model uses ρ , 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. We estimate ρ as the probability a fan link connecting the first to second user of a randomly selected pair of users, corresponding to the average number of fans per user divided by the number of active users U .

4.4 Visibility to submitter's fans

Because stories are always visible to fans and we know the number of fans of the story's submitter, the model behavior (Eq. (1) and (4)) can be solved without reference to the other parts of the model. We have $P_S = 1$ when the story

is on the front page and $P_S = c_S < 1$, reflecting users' preference for front page stories. Thus, these equations have two story-independent parameters, i.e., the rate users visit Digg (ω) and the probability users view upcoming stories submitted by their friends (c_S), and two story-dependent parameters, i.e., the interestingness (r_S) and number of fans of the submitter (S_0). S_0 is given in our data, while we estimate the other parameters from the data, i.e., votes by fans of the stories' submitters.

4.5 Visibility to non-fans

In our model, story location affects visibility only for non-fan voters since fans of prior voters see the story via the friends interface. Thus we use just the non-fan votes to estimate visibility parameters, via maximum likelihood. Specifically, we note a story typically receives only a few dozen votes before promotion, mostly from fans. With the value of ρ , estimated as described above, Eq. (6) gives $N(t) \approx U$ up to a few hours after promotion. Over this time period, Eq. (3) simplifies to $dN/dt \approx \omega U r_N P_N$ with P_N depending on story location on the recency and popularity lists. r_N is constant for a given story, so P_N determines the time variation in the voting rate by non-fans.

For front page stories, in our model $P_N = P_{\text{visibility}}(t, v)$ from Eq. (9), which has three parameters: μ and λ characterizing the browsing behavior for the recency and popularity lists, and the probability to find the story by other methods, β . We estimate these parameters by maximizing the likelihood of observing the non-fan front page votes according to the model, as described above for estimating a Poisson process with a time-dependent rate in Eq. (13). This estimation also determines r_N for each story.

For upcoming stories, we take $P_N = c_N P_{\text{visibility}}(t, v)$, giving a single additional parameter, c_N , to estimate, since we assume browsing behavior on the upcoming pages is the same as for front pages. This assumption has little effect on the model behavior because of the large number of submissions and relatively few non-fan votes for upcoming stories. Specifically, a submitted story remains near the front of the recency list for only about a minute after submission and stories reaching the front of the popularity list (due to having many votes) are soon promoted to the front page. Thus moderate variations in how deeply users browse the upcoming recency or popularity lists (i.e., the values for μ and λ) have little effect on the non-fan votes. Instead, the relatively few non-fan upcoming votes arise mainly through users finding the story by other means (e.g., a link from the submitter's blog during the few hours between a story's submission and its promotion). That is, in most cases $P_{\text{visibility}} \approx \beta$ for upcoming stories. Thus $P_N = c_N P_{\text{visibility}}(t, v) \approx c_N \beta$ and any difference between β for upcoming and front page stories would merely be reflected as a change in the value of c_N . This parameter is readily estimated using Eq. (13) with the upcoming non-fan votes.

parameter	value
average rate each user visits Digg	$\omega = 0.16 \pm 0.01$ /hr
number of active users	$U = 248,000 \pm 3000$
page view distribution	$\mu = 0.92 \pm 0.04$
	$\lambda = 0.9 \pm 0.1$
visibility by other methods	$\beta = 0.05 \pm 0.01$
probability a user is a voter's fan	$\rho = 1.7 \times 10^{-5}$
upcoming stories location	$k_u = 59.8$ pages/hr
front page location	$k_f = 0.31$ pages/hr
fraction viewing upcoming pages	
submitter fans	$c_S = 0.57 \pm 0.03$
other fans	$c_F = 0.10 \pm 0.01$
non-fans	$c_N = 0.11 \pm 0.01$
story specific parameters	
interestingness to submitter fans	r_S
interestingness to other fans	r_F
interestingness to non-fans	r_N
number of submitter's fans	S_0
promotion time	$T_{\text{promotion}}$

Table 1 Model parameters, with times in Digg hours.

4.6 Visibility to other fans

From Eq. (2), dv_F/dt changes abruptly when the story is promoted since P_F changes from c_F to 1 upon promotion. Thus we estimate c_F by the change in voting rate by fans other than those of the submitter by comparing the votes a story receives one hour before promotion and the votes received during the hour after promotion.

With all story-independent parameters estimated, we can then solve the full model for a story to determine dv_F/dt as a function of time. This gives the expected rate of other fan votes as a function of time. We determine r_F for the story as the value maximizing the log-likelihood (Eq. (13)) for the other fan votes the story receives.

4.7 Summary

Table 1 lists the estimated parameters. Note that all of these parameters, except the three story interestingness parameters r_S , r_F and r_N , are either known (e.g., the number of submitter's fans) or estimated from data. The interestingness parameters are estimated for each story from the votes.

vote type	μ	σ
submitter fan	-3.5 ± 0.2	0.8 ± 0.1
other fan	-2.3 ± 0.1	0.3 ± 0.1
non-fan	-6.3 ± 0.1	0.6 ± 0.1

Table 2 Parameters for lognormal distribution of interestingness.

5 Results

Fig. 1 compares the solution of the rate equations with the actual votes for one story. This illustrates the model captures the main qualitative features of the vote dynamics: an abrupt jump in votes after promotion followed by a slowing of the voting rate.

Fig. 6 shows how visibility estimated by our model (indicated by color) compares with the distribution of front page votes. Many votes occur when the story is recently promoted (so near the top of the recency list) or has received many votes within 24 hours after promotion (so near the top of the popularity list). Reassuringly, our model predicts higher visibility for stories in these positions on the lists.

5.1 Interestingness for fans and non-fans

We use the model to evaluate systematic differences in story interestingness between fans and non-fans. The estimated r values for the stories in our data set show the promoted stories have a wide range of interestingness to users, as shown in Fig. 9, along with fits to lognormal distributions. The figure shows r_N values tend to be much smaller than the interestingness for fans, as also seen in an earlier study with a smaller data set from 2006 [12]. The r -values are weakly correlated, with Spearman rank correlation between r_S and r_F of 0.20, between r_S and r_N of 0.22, and between r_F and r_N of 0.13. Moreover, there is a large range in the ratio of interestingness to fans and non-fans, suggesting stories with particularly large ratios are mainly of niche interest in the user community.

Table 2 summarizes the lognormal distribution parameters. A bootstrap test [7] based on the Kolmogorov-Smirnov (KS) statistic shows the estimated r -values are consistent with this distribution (p -value 0.11, 0.14 and 0.05 for the three cases). This test and the others reported in this paper account for the fact that we fit the distribution parameters to the data [5].

The relationship between interestingness for fans and other users indicates a considerable variation in how widely stories appeal to the general user community. Moreover, we find other fans have somewhat higher interest in stories than submitter fans, i.e., r_F tends to be larger than r_S (especially for stories of relatively little interest to the submitter’s fans). Since we have $c_S > c_F$

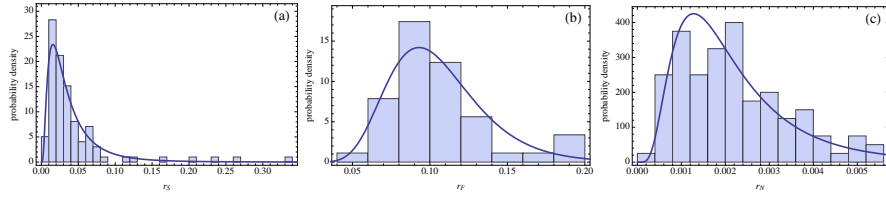


Fig. 9 Distribution of interestingness for (a) submitter fans, (b) other fans, and (c) non-fans. The curves are lognormal fits to the values. The axes scales differ among the plots.

(Table 1), we find submitter fans are more likely to view the story while upcoming, but less likely to vote for it, compared with other fans. This suggests people favor the submitter as a source of stories to read, while the fact that a friend, not the submitter, voted for the story makes it more likely the user will vote for the story. Alternatively, people who focus on submissions tend to be more likely to visit upcoming pages to get the latest news, while those learning about the story as a fan of some other voter are more likely to vote on it. Identifying these possibilities illustrates how models can suggest subgroups of behaviors in social media for future investigation.

5.2 Predicting popularity from early votes

In this section we investigate the use of the stochastic model to predict popularity of Digg stories. We study in detail the 89 of the 100 stories in the calibration data set that were promoted within 24 hours of submission. We focus on these stories because most stories are promoted within 24 hours of submission (if they are ever promoted) and this restriction simplifies the model’s use of the “popular in last 24 hours” list by not requiring it to check for removal from the list if the story is still upcoming more than 24 hours after submission. Predicting popularity in social media from intrinsic properties of newly submitted content is difficult [25]. However, users’ early reactions provide some measure of predictability [13, 15, 18, 26]. The early votes on a story allow estimating its interestingness to fans and other users, thereby predicting how the story will accumulate additional votes. These predictions are for expected values and cannot account for the large variation due, for example, to a subsequent vote by a highly connected user which leads to a much larger number of votes.

We can improve predictions from early votes by using the lognormal distributions of r -values, shown in Fig. 9, as the prior probability to combine with the likelihood from the observations according to Bayes theorem. Specifically, instead of maximizing the likelihood of the observed votes, $P(r|\text{votes})$, as discussed above, this approach maximizes the posterior probability, which is proportional to $P(r|\text{votes})P_{\text{prior}}(r)$ where P_{prior} is taken to be the lognormal distribution $P_{\text{lognormal}}$ in Eq. (14) with parameters from the fits shown in Fig. 9.

For a prediction at time T , we use the votes up to time T to estimate the r values by finding the values that maximize

$$L = \log(P(r_S, r_F, r_N | \text{votes})) + \log(P_{\text{prior}}(r_S)P_{\text{prior}}(r_F)P_{\text{prior}}(r_N)) \quad (15)$$

We then solve the model starting at time T and use the values from that solution as the predictions at later times. Solving the model equations starting at time T requires initial values, i.e., the number of votes $v_S(T)$, $v_F(T)$, $v_N(T)$ and the size of the user groups who have not yet seen the story: $S(T)$, $F(T)$, $N(T)$. The numbers of votes, also used to estimate the r values, is available in our data. However, the sizes of the user groups is not available. Instead, we estimate these values from the voting *rates* and the estimated r values. For instance, Eq. (1) gives

$$S(T) = \frac{1}{wr_S P_S} \frac{dv_S}{dt} \quad (16)$$

We estimate the voting rates from the number of votes in the 15 minutes prior to time T , except if there are fewer than five votes in this time we extend the time interval to include the five previous votes. For simplicity, to avoid treating the discontinuity in visibility at promotion, we based this estimate on front page votes when T is after the promotion time.

We focus on behavior after promotion. This is particularly relevant in a web site such as Digg where early estimates of likely eventual popularity could suggest content to highlight to users. Fig. 10 compares predicted to actual votes for one story 24 hours after promotion. Votes from submission to promotion are used to estimate r values for the three classes of users. The model solutions are then extrapolated from the time of these estimates, i.e., the story's promotion time, to $t = 24$ Digg hours after promotion. The model quantitatively, as well as qualitatively reproduces the observed votes for this story.

Generalizing from this example for a single story, Fig. 11 shows the prediction errors when using the model to predict the number of each type of vote the story receives 24 Digg hours after promotion, based on estimating r -values from early votes observed up to time T for various times T at and shortly after promotion. For context of the size of these errors, Fig. 5 shows the range of number of votes the stories have at the times of these predictions, i.e., 24 hours after promotion.

As expected, errors generally decrease when predictions are made later. Of more interest is the difference among the type of votes, particularly for votes from other fans. Early votes are mainly from submitter's fans and non-fans, so the ability to predict differences in behavior for those groups based on early votes could be useful in quickly distinguishing stories likely to be of broad interest to the user community from niche interests.

Overall, the model reasonably predicts votes from submitter's fans and non-fans, but is much less accurate for votes from other fans. One reason for this difference is the relatively small number of other fan votes while a story is upcoming. Specifically, the pool of other fans F starts at zero. Only a vote by

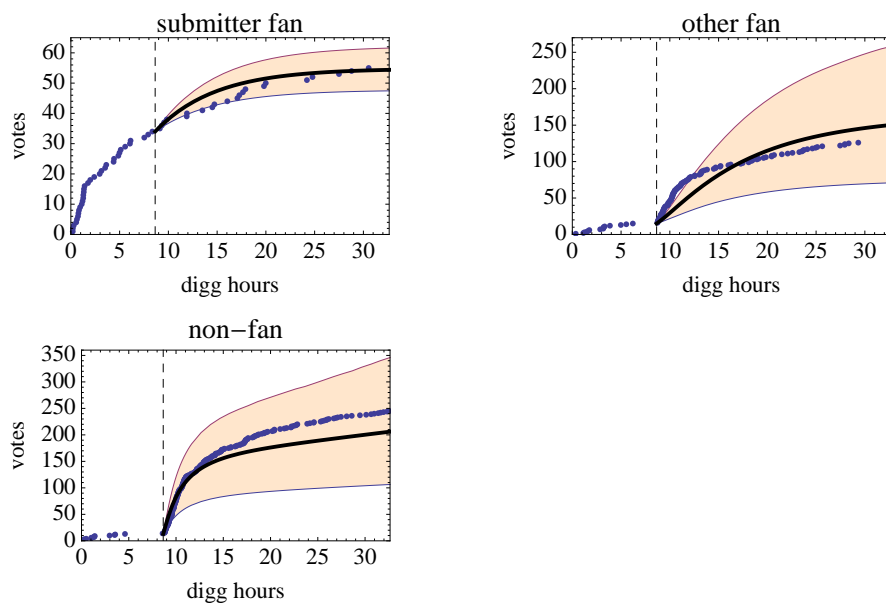


Fig. 10 Predictions compared to actual votes (dots) for each type of user for one story. The figure shows predictions made at promotion (black line) and the growth in the 95% confidence interval of the prediction up to 24 hours after promotion. The dashed vertical line shows the story's promotion time.

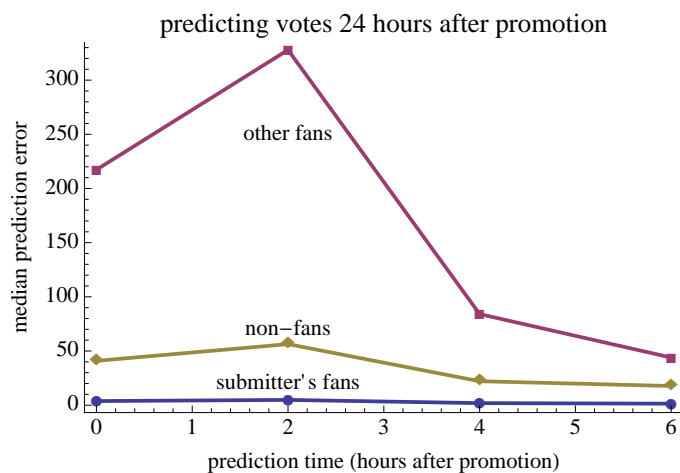


Fig. 11 Median error between predicted and observed votes 24 Digg hours after promotion for predictions made 0, 2, 4 and 6 Digg hours after promotion.

a non-fan can increase F , and upcoming stories have low visibility to non-fan voters. Even after a pool of other fans becomes available, it takes some time for those users to return to Digg. Thus there are relatively few early other fan votes, leading to poor estimates for r_F values. Moreover, the relatively small pool of other fans means a single early voter with many fans can significantly

$T - T_{\text{promotion}}$	<i>correlation</i>		
	submitter fan	other fan	non-fan
0	0.88	0.26	0.51
2	0.95	0.59	0.85
4	0.98	0.69	0.92
6	0.99	0.75	0.94

Table 3 Spearman rank correlation between predicted and observed number of each type of votes 24 Digg hours after promotion, for predictions made at various times T after promotion (measured in Digg hours).

$T - T_{\text{promotion}}$	<i>classification error</i>		
	submitter fan	other fan	non-fan
0	0.19	0.40	0.29
2	0.11	0.48	0.17
4	0.10	0.37	0.15
6	0.06	0.28	0.12

Table 4 Classification errors on whether a story receives more than the median number of votes from each type of voter received by 24 Digg hours after promotion, for predictions made at various times T after promotion (measured in Digg hours).

change F away from its average value as used in the model. This combination of factors leads to the relatively large errors in predicting the other fan votes. As a direction for future work, this observation suggests predictions would benefit from including measurements of the social network of the voters to determine the value of F at the time of prediction rather than using an estimate based on the model.

Another view of prediction quality is how well the model predicts the rank ordering of stories, i.e., whether the story is likely to be relatively popular. We measure this with the Spearman rank correlation between the model’s prediction and the observed number of votes 24 Digg hours after promotion, as shown in Table 3. Even for other fan votes, where the absolute prediction error is relatively large, the predicted values give a good indication of the relative rank of the stories.

Predicting whether a story will attract a large number of votes, rather than the precise number of votes, is a key issue for web sites such as Digg. Such predictions form the basis of using crowd sourcing to select a subset of submitted content to highlight [18]. As an example of this distinction, we predict whether a story will receive more than the median number of votes of each type of user based on votes received up to various times. This amounts to a binary classification task. Table 4 compares predictions made at different times. The classification error rate is the fraction of stories for which prediction of whether the story receives more than the median number of votes differs from the actual value. The model generally overestimates the number of votes from other fans and predicts it to be above the median value, as indicated by error rates close to 0.5.

5.3 Confidence intervals

We can use the model to estimate how well it will predict future votes. For a given set of parameter values, prediction variability comes from differences in estimated r values. If the r value is poorly determined, predictions based on this maximum value will be unreliable.

To quantify this behavior, we numerically evaluate the second derivative matrix D of the log-likelihood combined with the priors based on votes on the story up to time T , $L(r)$ given in Eq. (15), at the maximum $r = r_{\max}$, where $r = (r_S, r_F, r_N)$. This gives

$$L(r) = L(r_{\max}) + \frac{1}{2}(r - r_{\max})D(r - r_{\max}) \quad (17)$$

to second order in $|r - r_{\max}|$. To this order of expansion, the likelihood is

$$\exp(-(r - r_{\max})D(r - r_{\max})/2) \quad (18)$$

This corresponds to a multivariate normal distribution for r with mean r_{\max} and covariance matrix $-D^{-1}$. Since we are expanding around a maximum, the 2nd derivative matrix is negative definite so this gives a well-defined normal distribution, i.e., with a positive definite covariance matrix. This covariance includes both individual variances in the values of r_S , r_F and r_N and correlations among their variations around the maximum.

If $L(r)$ is a fairly flat function of r around the maximum then maximum likelihood poorly constraints the values, corresponding to large variances in the normal distribution. Conversely, if $L(r)$ is sharply peaked, the distribution will be narrow.

We apply this observation to estimate confidence intervals for the predictions. We first numerically evaluate the second derivative matrix D at the maximum. We then generate random samples of r from the multivariate normal distribution. For each of these samples, we solve the model starting from the time T to any desired time for predicting the votes, e.g., 24 hours after promotion. After collecting these predictions from many samples, we use quantiles of their ranges as the confidence intervals. In the examples presented here, we generate 1000 random samples and determine the 95% confidence interval from the variation in r values as the range between the 2.5% and 97.5% quantiles of these samples.

As one example, Fig. 10 shows how confidence intervals grow with time for predictions made from votes at the time a story is promoted. For multiple stories, Fig. 12 shows the relation between 95% confidence interval and prediction error 24 Digg hours after promotion, based on prediction made at the time of promotion. We see generally that large errors are associated with large confidence intervals, especially for the other fan votes where the model's prediction errors are largest. This indicates the confidence intervals, which are

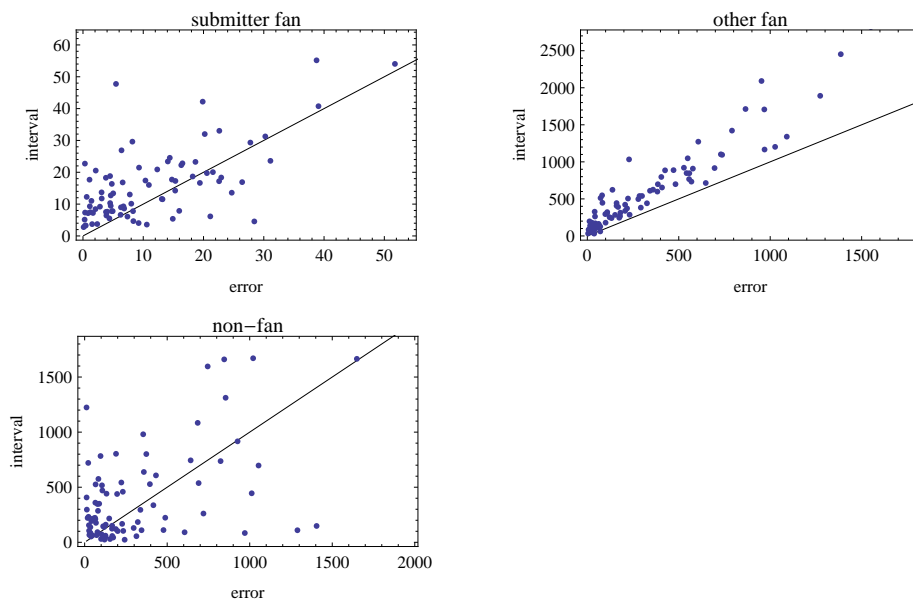


Fig. 12 Size of 95% confidence interval vs. prediction error 24 Digg hours after promotion for predictions based on votes up to each story's promotion time. The diagonal lines correspond to errors equal to the size of the confidence interval.

computed from the vote information available at the time of prediction, provide an indication of how well the model can predict votes. These scatterplots show some cases where the error is considerably larger than the confidence interval. As these cases are significantly more than $1/20$ of the points, they indicate additional sources of variation not accounted for by the variation in the estimated r values. This could be due, for instance, to votes by exceptionally well-connected users that significantly increase the story's visibility compared to the average value assumed with the model.

6 Related Work

Social dynamics can help explain and predict the popularity of online content. The broad distributions of popularity and user activity on many social media sites can arise from simple macroscopic dynamical rules [27]. 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 [28]. Stochastic models [17, 11] offer an alternative explanation for the vote distribution. Rather than novelty decay, they explain the votes distribution by the combination of variation in the stories' inherent interest to users and effects of user interface, specifically decay in visibility as the story moves to subsequent pages. Crane and Sornette [6] found that collective dynamics was linked to the inherent quality of videos on YouTube. From the number of votes re-

ceived by videos over time, they could separate high quality videos from junk videos. This study is similar in spirit to our own in exploiting the link between observed popularity and content quality. However, 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.

Statistically significant correlation between early and late popularity of content is found on Slashdot [15], Digg and YouTube [26]. Specifically, similar to our study, Szabo & Huberman [26] predicted long-term popularity of stories on Digg. Through large-scale statistical study of stories promoted to the front page, they were able to predict stories' popularity after 30 days based on its correlation with popularity one hour after promotion. Similarly, Lerman & Hogg [19] predicted popularity of stories based on their pre-promotion votes. We also quantitatively predict stories' future popularity, but unlike earlier works, we can also estimate confidence intervals of these predictions.

Previous works found social networks to be an important component to information diffusion. Niche interest content spreads mainly along social links in Second Life [2], as well as on Digg [18], and does not end up not becoming very popular out-of-network. Aral et al. [1] found that social links between like-minded people, rather than causal influence, explained much of information diffusion observed on a network. Our modeling approach allows us to systematically distinguish users who are linked to those who are not linked and study diffusion separately for each class of people.

7 Discussion

Highlighting friends' contributions is a common feature of social media sites, including Digg. To evaluate the effects of this behavior, we explicitly include votes from submitter's fans, other fans and non-fans in our model, while separating the effects of differences in visibility and interestingness between these groups of users. This identifies that submitter's fans are, on average, far more likely to find the story interesting. Our model adjusts for the higher visibility of stories to fans, thereby identifying the increased attention from fans is not just due to the increased visibility. Identifying stories of particularly high interest to fans could be a useful guide for highlighting stories in the friends interface, i.e., emphasizing those with relatively large interestingness to friends as reflected in the early votes. Moreover, this information could be useful to recommend new fans to users, based on visibility-adjusted similarity in voting rather than, as commonly done in collaborative filtering [16], just using the raw score of similar votes. This could be particularly important for users with relatively infrequent votes, where variations due to how visible a story is could significantly affect the similarity of the vote pattern with that of other users.

For more precise estimates, the web site could track the fraction of users seeing the story that vote for it, thereby directly estimating interestingness and accounting for the large variability in number of fans among the voters, in contrast to our model which used an average value. Exploiting such details of user behavior becomes more important as the complexity of the web site interface increases, offering many ways for users to locate content. Recording which method leads each user to find the story can aid in identifying any systematic differences in interests among those users, generalizing our study which distinguished users who could or could not find the story in their friends interface.

We find a wide range of interestingness ratios between fans and non-fans. This explains prior observations of the effect of relatively high votes from fans on indicating popularity to the general user population, and also suggests stories that are of niche interest to the fans rather than the general user population. Our assumption that fans of prior voters easily see the story is reasonable for users with relatively few fans, so only a few stories will appear in their friends interface. When this is not the case, visibility of a story would decrease when many newer stories appear on the friends interface. This possibility could be included in the model using the “law of surfing” fan votes based on the number of stories appearing in each users’ friends interface.

For prediction, we find the largest errors with votes from other fans. This likely arises from the relatively small number of such votes, especially while the story is upcoming. In that case, the large variation in number of fans per user can have a dramatic effect not accounted for in the model: if a user with many fans is an early voter for the story, the number of fans who have not yet seen the story, i.e., F , will increase significantly, thereby giving a larger number of potential other fan voters than accounted for in the model. Subsequent votes by these users will appear to indicate a large interest by other fans (i.e., large value of r_F) leading to a prediction for many such votes. This suggests the main source of the prediction error arises from the long-tail distribution of fans per user, which the model treats as a single average value based on the parameter ρ . We could test this possibility by collecting additional data on the actual fans of each voter, thereby including the observed value of $F(t)$ at the time of prediction when estimating r -values. In cases where $F(t)$ is particularly large, e.g., due to an early vote by a user with many fans, this will result in a smaller estimated value for r_F and hence smaller predicted number of other fan votes.

One use of models is to suggest improved designs for user-contributory web sites. Our results suggest it may be useful to keep popular stories visible longer for users who return to Digg less often – giving them more of a chance to see the popular stories before they lose visibility. This would be a fine tuned version of “popular stories” pages, adjusted for each user’s activity rate. That is, instead of showing stories in order of recency only, selectively move less popular stories down the page (once there are enough votes to determine popularity), thereby

leaving the more popular ones nearer the top of the list for users who come back to Digg less often.

We examined behavior over a relatively short time (e.g., up to a day after promotion). Over longer times, additional factors could become significant, particularly a decrease in the interestingness as news stories submitted to Digg become “old news” [28].

Modeling visibility depends on how the web site user interface exposes content. This highlights a challenge for modeling social media: continual changes to the user interface can alter how visibility changes for newly submitted content. Thus accurate models require not only data on user behavior but also sufficient details of the user interface at the time of the data to determine what properties of the content determine visibility.

The lognormal distribution of interestingness seen here and in other web sites [13], could be useful as a prior distribution for estimating interestingness from early behavior on web sites. In particular, for early estimates where there are only a few votes, observing zero votes from fans would lead to the maximum likelihood estimate $r_F = 0$ while using the lognormal prior would estimate a small, but nonzero, value. The use of such priors will be more important as models make finer distinctions among groups of users, e.g., distinguishing those who find the content in many different ways as provided by more complex interfaces. In such cases, many groups will not be represented among the early reaction to new content, leading to maximum likelihood estimates of zero interest by those groups.

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

1. Sinan Aral, Lev Muchnik, and Arun Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544–21549, December 2009.
2. Eytan Bakshy, Brian Karrer, and Lada A. Adamic. Social influence and the diffusion of user-created content. In *Proc. of the 10th ACM Conf. on Electronic Commerce (EC09)*, pages 325–334, NY, 2009. ACM.
3. M. G. Bulmer. On fitting the Poisson lognormal distribution to species-abundance data. *Biometrics*, 30:101–110, 1974.
4. Meeyoung Cha, Hamed Haddadiy, Fabricio Benevenutoz, and Krishna P. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. In *Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM)*, 2010.
5. Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51:661–703, 2009.
6. Riley Crane and Didier Sornette. Viral, quality, and junk videos on YouTube: Separating content from noise in an information-rich environment. In K. Lerman et al., editors, *Proc. of the AAAI Symposium on Social Information Processing*, pages 18–20, 2008.

7. Bradley Efron. Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7:1–26, 1979.
8. Richard Haberman. *Mathematical Models: Mechanical Vibrations, Population Dynamics, and Traffic Flow*. Classics in Applied Mathematics. Society for Industrial Mathematics, 1987.
9. Herbert W. Hethcote. The Mathematics of Infectious Diseases. *SIAM REVIEW*, 42(4):599–653, 2000.
10. Joseph M. Hilbe. *Negative Binomial Regression*. Cambridge Univ. Press, 2008.
11. Tad Hogg and Kristina Lerman. Stochastic models of user-contributory web sites. In *Proc. of the Third International Conference on Weblogs and Social Media (ICWSM2009)*, pages 50–57. AAAI, 2009.
12. Tad Hogg and Kristina Lerman. Social dynamics of Digg. In *Proc. of the Fourth International Conference on Weblogs and Social Media (ICWSM2010)*, pages 247–250, Menlo Park, CA, 2010. AAAI.
13. Tad Hogg and Gabor Szabo. Diversity of user activity and content quality in online communities. In *Proc. of the Third International Conference on Weblogs and Social Media (ICWSM2009)*, pages 58–65. AAAI, 2009.
14. Bernardo A. Huberman, Peter L. T. Pirolli, James E. Pitkow, and Rajan M. Lukose. Strong regularities in World Wide Web surfing. *Science*, 280:95–97, 1998.
15. A. Kaltenbrunner, V. Gomez, and V. Lopez. Description and prediction of slashdot activity. In *Proc. 5th Latin American Web Congress (LA-WEB 2007)*, 2007.
16. J. A. Konstan, B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon, and J. Riedl. GroupLens: Applying collaborative filtering to Usenet news. *Communications of the ACM*, 40(3):77–87, 1997.
17. K. Lerman. Social information processing in social news aggregation. *IEEE Internet Computing: special issue on Social Search*, 11(6):16–28, 2007.
18. K. Lerman and A. Galstyan. Analysis of social voting patterns on Digg. In *Proceedings of the 1st ACM SIGCOMM Workshop on Online Social Networks*, 2008.
19. Kristina Lerman and Tad Hogg. Using a model of social dynamics to predict popularity of news. In *Proc. of the 19th Intl. World Wide Web Conference (WWW2010)*, pages 621–630, NY, 2010. ACM.
20. Kristina Lerman and Tad Hogg. Using stochastic models to describe and predict social dynamics of web users. *Submitted to ACM Transactions on Intelligent Systems and Technology*, 2011.
21. Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne Vanbriesen, and Natalie Glance. Cost-effective outbreak detection in networks. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 420–429, New York, NY, USA, 2007. ACM.
22. Guthrie Miller. Statistical modelling of Poisson/log-normal data. *Radiation Protection Dosimetry*, 124:155–163, 2007.
23. Tye Rattenbury, Nathaniel Good, and Mor Naaman. Towards automatic extraction of event and place semantics from Flickr tags. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 103–110, New York, NY, USA, 2007. ACM.
24. William J. Reed and Murray Jorgensen. The double Pareto-lognormal distribution: A new parametric model for size distributions. *Communications in Statistics: Theory and Methods*, 33:1733–1753, 2004.
25. M.J. Salganik, P.S. Dodds, and D.J. Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311:854, 2006.
26. Gabor Szabo and Bernardo A. Huberman. Predicting the popularity of online content. *Social Science Research Network Working Paper Series*, November 2008.
27. Dennis M. Wilkinson. Strong regularities in online peer production. In *EC '08: Proceedings of the 9th ACM conference on Electronic commerce*, pages 302–309, New York, NY, USA, 2008. ACM.
28. Fang Wu and Bernardo A. Huberman. Novelty and collective attention. *Proceedings of the National Academy of Sciences*, 104(45):17599–17601, November 2007.