

DOI:10.1145/1787234.1787254

Early patterns of Digg diggs and YouTube views reflect long-term user interest.

BY GABOR SZABO AND BERNARDO A. HUBERMAN

Predicting the Popularity of Online Content

THE EASE OF producing online content highlights the problem of predicting how much attention any of it will ultimately receive. Research shows that user attention⁹ is allocated in a rather asymmetric way, with most content getting only some views and downloads, whereas a few receive the most attention. While it is possible to predict the distribution of attention over many items, it is notably difficult to predict the amount that will be devoted over time to any given item. We solve this problem here, illustrating our approach with data collected from the portals Digg (<http://digg.com>) and YouTube (<http://youtube.com>), two well-known examples of popular content-sharing-and-filtering services.

The ubiquity of Web 2.0 services has transformed the landscape of online content consumption. With the Web, content producers can reach an audience in numbers inconceivable through conventional

channels. Examples of services that have made the exchange between producer and consumer possible on a global scale include video, photo, and music sharing, blogs, wikis, social bookmarking, collaborative portals, and news aggregators, whereby content is submitted, perused, rated, and discussed by the user community.

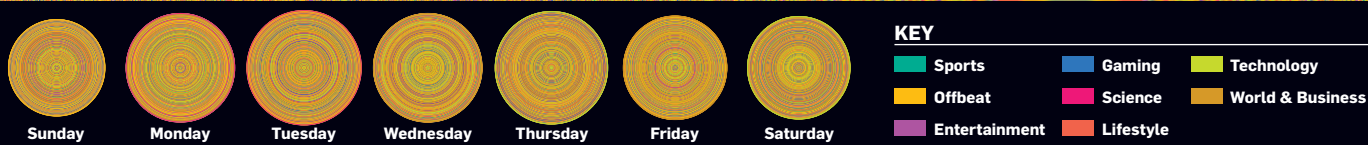
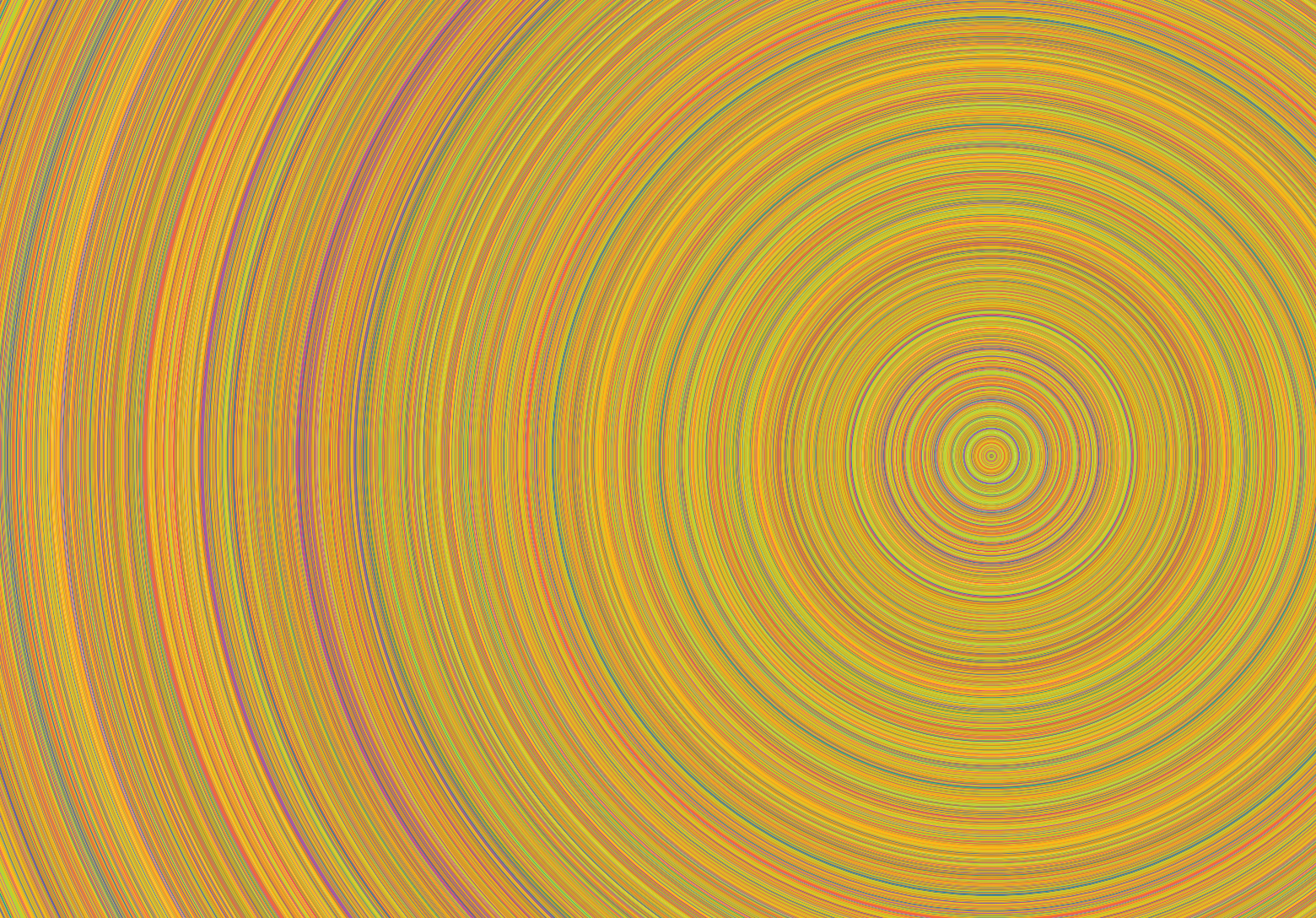
Portals often rank and categorize content based on past popularity and user appeal, especially for aggregators, where the “wisdom of the crowd” provides collaborative filtering to select submissions favored by as many visitors as possible. Digg is an example, with users submitting links to and short descriptions of content they have found on the Web and others voting on them if they find them interesting. The articles attracting the most votes are exhibited on the site’s premiere sections under headings like “recently popular submissions” and “most popular of the day.” This placement results in a positive feedback mechanism leading to rich-get-richer vote accrual for the very popular items, though the pattern pertains to only a small fraction of the submissions that rise to the top.

Besides Digg, anyone with Internet access can watch YouTube videos, reply to them through their own videos, and leave comments. The way the online ecosystem has developed around YouTube videos is impressive by any standard, and videos that draw millions of viewers are prominently displayed on the site, like stories on Digg.

Content providers, Web hosts, and advertisers all would like to be able to predict how many views and downloads individual items might generate on a given Web site. For example in advertis-

» key insights

- Site administrators, advertisers, and providers would all find it useful to be able to predict content popularity.
- Prediction is possible due to the extreme regularity with which user attention focuses on content.
- Early patterns of access indicate long-term popularity of content.



Chris Harrison's Digg Rings visualization plots the top 10 most-dugg stories by days of the week May 24, 2007 to May 23, 2008 (bottom) and all stories (close up) Dec. 1, 2004 to May 23, 2008 (top) rendered as a series of tree-ring-like visualizations moving outward in time.

ing, if popularity count is tied directly to ad revenue (such as with ads shown with YouTube videos), revenue might fairly accurately be estimated ahead of time if all parties know how many views the video is likely to attract. Moreover, in content-distribution networks, the computational requirements for bandwidth-intensive new content may be determined early on if the hosting site is able to extrapolate the number of requests the content is likely to get by observing patterns of access from the moment it was first posted.

Digg allows users to submit links to news, images, and videos they find on the Web and think will interest the site's general audience. Based on data we collected from Digg in the second half of 2007, 90.5% of all uploads were links to news, 9.2% to videos, and only

0.3% to images. Submitted content is placed by the submitters on Digg in the so-called "upcoming" section, one click from the site's main page. Links to content are provided, along with surrogates to the submission (a short description for news, a thumbnail image for images and videos) intended to entice readers to peruse the content. Digg functions as a massive collaborative filtering tool to select and share the most popular content in the user community; registered users thus digg submissions they find interesting. Digging increases the digg count of the submission one digg at a time, and submissions that get enough diggs in a certain amount of time in the "upcoming" section are shown on the Digg front page, or, per Digg terminology, "promoted." Promotion is a considerable source

of pride in the Digg community and a main motivator for repeat submitters. The exact algorithm for promotion is not made public to thwart gaming but is thought to give preference to upcoming submissions that accumulate diggs quickly from diverse neighborhoods in the Digg social network,⁷ thus modulating the influence of very popular submitters with hundreds of followers. Digg's social-networking feature lets users place watch lists on other users by becoming their fans. Fans are shown updates on which submissions are digg by these users; the social network therefore plays a major role in making upcoming submissions more visible to a larger number of users. Here, we consider only stories that were promoted, since we were interested in submissions to which many users had access.

We used the Digg application programming interface (<http://apidoc.digg.com/>)⁴ to retrieve all diggs made by registered users from July 1, 2007 to December 18, 2007. This data set included approximately 29 million diggs by 560,000 users on approximately 2.7 million submissions, a number including all past submissions receiving any digg, not only the submissions during the six months. The number of submissions was about 1.3 million, of which about 94,000 (7.1%) were promoted to the front page.

YouTube is the apex of the Web’s user-created video-sharing portals, with (as of 2008) 65,000 new videos uploaded and 100 million viewed daily, implying that 60% of all online videos were watched through YouTube.^{3,6} It was also the third most frequently accessed site on the Web, based on traffic rank.¹ Beginning April 21, 2008, we collected view-count time series on 7,146 selected videos daily in the portal’s “recently added” section, carrying out data collection for the next 30 days. Apart from the list of “most recently

added” videos, it also offered listings based on such YouTube-defined selection criteria as “featured,” “most discussed,” and “most viewed.” We chose the “most recently uploaded” list to give us an unbiased sample of all videos submitted to the site or complete history of the view counts for each video during its lifetime. YouTube’s API (<http://code.google.com/apis/youtube/overview.html>)¹⁰ provided programmatic access to several video statistics, with view count at a given time being one of them.

However, due to the fact that the view-count field of a video did not appear to have been updated more often than once a day by YouTube, we were able to calculate only a good approximation of the number of daily views. Worth noting is that while the overwhelming majority of video views was initiated from the YouTube Web site itself, videos might have been linked from external sources as well, appearing as embedded objects on the referring page; while 50% of all videos in 2007 were thought to be linked externally, only about 3% of the views came from these links.²

Popularity Growth

By “popularity” we mean number of votes (diggs) a story collected on Digg and number of views a video received on YouTube, respectively. Figure 1 reflects the dynamics of content popularity growth on both portals, showing the average normalized popularity for all submissions over time; we first determined the popularity of each individual submission at the end of the 30th day following its submission, dividing their popularity values before that time by this final number. For each submission, we obtained a time series of popularities that monotonically increased from 0 (at submission time) to 1 at day 30. By thus eliminating the prevailing differences in content-specific interestingness among the submissions (one submission might get only a few views over its lifetime, while another gets thousands or even millions), we averaged overall submissions of the normalized popularities.

An important difference between the two portals is that while Digg stories saturate fairly quickly (about a day) to their respective reference populari-

Figure 1. Average normalized popularity of submissions to Digg and YouTube by individual popularity at day 30. The inset is the same measurement for the first 48 digg hours of Digg submissions.

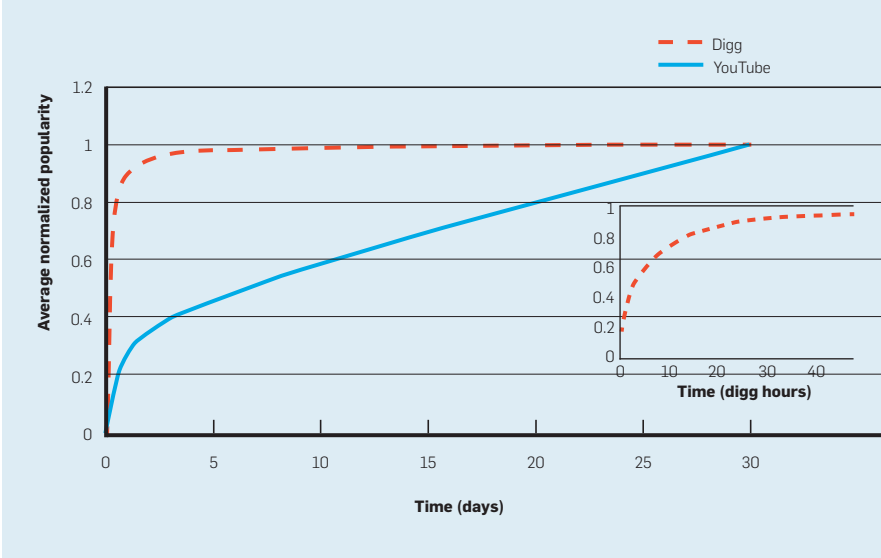
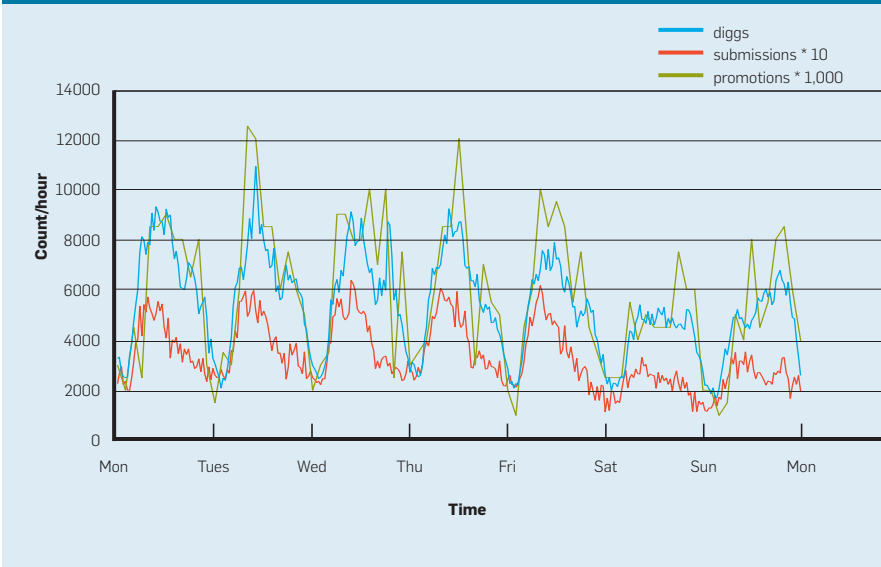


Figure 2. Daily and weekly cycles in the hourly rates of digging activity, story submissions, and story promotions, respectively. To match the different scales, we multiplied the rates for submissions by 10 and the rates of promotion by 1,000. The horizontal axis represents the week August 6, 2007 (Monday)–August 12, 2007 (Sunday). The tick marks are midnight on the respective day, Pacific Standard Time.



ties, YouTube videos keep attracting views throughout their lifetimes. The rate videos attract views may naturally differ among videos, with the less-popular likely marking a slower pace over a longer time.

These two notably different user-popularity patterns are a consequence of how users react to content on the two portals. On Digg, articles quickly become obsolete, since they often link to breaking news, fleeting Internet fads, or technology-related themes with a naturally limited time for user appeal. However, videos on YouTube are mostly found through search, since, with the sheer number of videos constantly being uploaded, it is not possible to match Digg's way of giving each promoted story general exposure on a front page. The quicker initial rise of video view counts can be explained through the videos' exposure in YouTube's "recently added" section, but after leaving it, the only way to find them is through keyword search or when displayed as related videos next to another video being watched.

The short fad-like popularity life cycle of Digg stories (a day or less) suggests that if overall user activity on Digg depends on time of day, a story's popularity may grow more slowly when fewer visitors are on the site and increase more quickly at peak periods. For YouTube, this effect is less relevant, since video views are spread over more time, as in Figure 1. Figure 2 outlines the hourly rates of user digging, story submitting, and upcoming Digg story promotions as a function of time for one week, beginning August 6, 2007. The difference in rates may be as much as threefold; weekends showed less activity, and weekdays appeared to involve about 50% more activity than weekends. It was also reasonable to assume that besides daily and weekly cycles, such activity also involved seasonal variations. Moreover, in 2007, Digg users were mostly located in the UTC-5 to UTC-8 time zones (the Western hemisphere).

Depending on the time of day a submission was made to the portal, stories differed greatly in the number of initial diggs they received. As we expected, stories submitted during less-active periods of the day accrued fewer diggs in the first few hours than stories submitted during peak hours. This was a natural consequence of suppressed

Figure 3. Correlation of digg counts on the 17,097 promoted stories in the data set older than 30 days. A k-means clustering separates 89% of the stories into an upper cluster; the other stories are a lighter shade of blue. The bold line indicates a linear fit with slope 1 on the upper cluster, with a prefactor of 5.92 (Pearson correlation coefficient of 0.90).

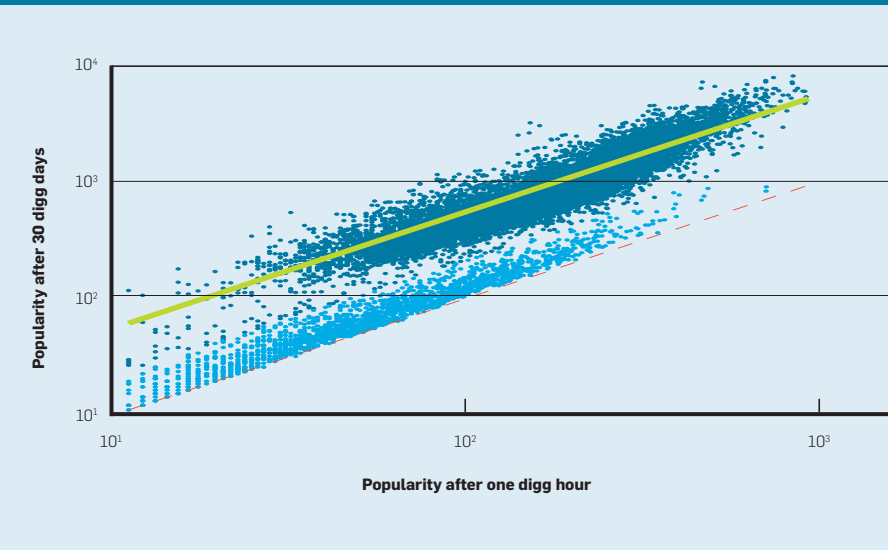
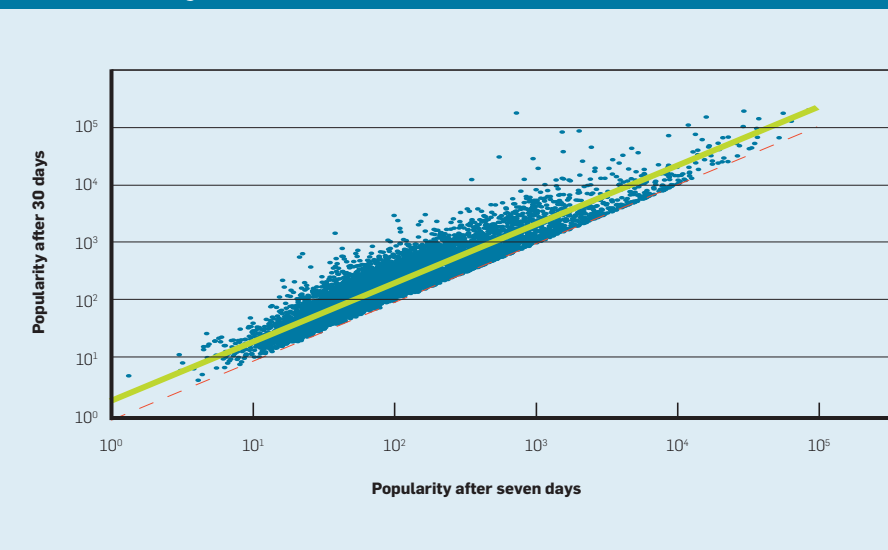


Figure 4. Popularity of videos on the 30th day after upload vs. popularity after seven days. The bold line with gradient 1 is fit to the data.



digging activity at night but might have initially penalized interesting stories that were otherwise likely to be popular. For instance, an average story promoted at 12 P.M. received approximately 400 diggs in the first two hours and only 200 diggs if promoted at 12 A.M. That is, based on observations made after only a few hours after a story was promoted, a portal could misinterpret the story's relative interestingness if it did not correct for the variation in daily user-activity cycles.

Since digging activity varies by time, we introduce the notion of digg time measured not in seconds but in number of diggs users cast on promoted sto-

ries. We count diggs only on promoted stories because this section of the portal was our focus, and most diggs (72%) were to promoted stories anyway. The average number of diggs arriving at promoted stories during any hour day or night was 5,478 when calculated over the full six-month data-collection period; we define one digg-hour as the time it takes for so many new diggs to be cast. As discussed earlier, the time for this many diggs to arrive took about three times longer at night than during the day. This "detrending" allowed us to ignore the dependence of submission popularity on the time of day it was submitted. Thus, when we refer to

the age of a submission in digg hours at a given time t , we measure how many diggs were received on the portal between t and the promotion time of the story, divided by 5,478 diggs.


Similar hourly activity plots were not possible for YouTube in 2008, given that video view counts were provided by the API approximately only once a day, in contrast to all the diggs received by a Digg story. Moreover, we were able to capture only a fraction of the large amount of traffic the YouTube site handled by monitoring only the selected videos in our sample.

Predicting the Future


Here, we cover the process we used to model and predict the future popularity of individual content and measure the performance of the predictions: First, we performed a logarithmic transformation on the popularities of submissions. The transformed variables exhibit strong correlations between early and later time periods; on this scale, the naturally random fluctuations can be expressed as an additive noise term. We call reference time t_r , the time at which we intend to predict the popularity of a submission whose age with respect to its upload (promotion) time is t_r . By indicator time t_i we mean when in the life cycle of the submission we performed the prediction, or how long we can observe submission history in order to extrapolate for future popularity ($t_i < t_r$).

To help determine whether the popularity of submissions early on is a predictor of later popularity, see Figures 3 and 4, which show the popularity counts for submissions at the reference time $t_r = 30$ days both for Digg and YouTube vs. the popularity measured at the indicator times $t_i = 1$ digg hour and $t_i = 7$ days for the two portals, respectively. We measured the popularity of YouTube videos at the end of the seventh day, so the view counts at that time ranged from 10^1 to 10^4 , similar to Digg in this measurement. We logarithmically rescaled the horizontal and vertical axes in the figures due to the large variances present among the popularity of different submissions, which span three decades.

Observing the Digg data, we noted the popularity of about 11% of the stories (lighter blue in Figure 3) grew much



While Digg stories saturate fairly quickly (about a day) to their respective reference popularities, YouTube videos keep attracting views throughout their lifetimes.



more slowly than the popularity of the majority of submissions; by the end of the first hour of their lifetimes, they had received most of the diggs they will ever receive. The difference in popularity growth of the two clusters is perceivable until approximately the seventh digg hour, after which the separation vanishes due to digg counts of stories mostly saturating to their respective maximum values, as in Figure 1.

A Bayesian network analysis of submission features (day of the week/hour of the day of submission/promotion, category of submission, number of diggs in the upcoming phase) reveals no obvious reason for the presence of clustering; we assumed it arises when the Digg promotion algorithm misjudged the expected future popularity of stories, promoting stories from the “upcoming” phase unlikely to sustain user interest. Users lose interest much sooner in them than in stories in the upper cluster. We used k-means clustering, with $k = 2$ and cosine distance measure to separate the two clusters, as in Figure 3, and discarded the stories in the lower cluster.

Trends and randomness. Our in-depth analysis of the data found strong linear correlations between early and later times of the logarithmically transformed submission popularities, with correlation coefficients between early and later times exceeding 0.9. Such a strong correlation suggests the more popular submissions are at the beginning, the more popular they will also be later on. The connection can be described by a linear model:

$$\begin{aligned} \ln N(t_r) &= \ln [r(t_i, t_r)N(t_i)] + \xi(t_i, t_r) \\ &= \ln r(t_i, t_r) + \ln N(t_i) + \xi(t_i, t_r), \end{aligned}$$

where $N(t)$ is the popularity of a particular submission at time t ; $r(t_i, t_r)$ accounts for the linear relationship between the log-transformed popularities at different times; and ξ is a noise term (describing the randomness we observed in the data) that accounts for the natural variances in individual content dynamics beyond the expected trend in the model and is drawn from a fixed distribution with mean 0. It is important to note that the noise term is additive on the log-scale of popularities, justified by the fact that we found the strongest correlations on this

transformed scale. In light of Figures 3 and 4, the popularities at t_r also appear to be evenly distributed around the linear fit, taking only the upper cluster in Figure 3 and considering the natural cutoff $y = x$ in the data for YouTube. We also found that the noise term (given by the residuals after a linear fit in both the YouTube and the Digg data) is well described by a normal distribution on the logarithmic scale.

However, there is also an alternative explanation for the observed correlations: If we let t_i vary in the model just described we see that the popularity at the given time t_r should be described by the following formula, assuming the noise term in the model is distributed normally (t_0 is an early point in time after submission/promotion):

$$\ln N(t_r) = \ln N(t_0) + \sum_{\tau=t_0}^{t_r} \eta(\tau).$$

$\eta(\tau)$ is a random value drawn from an arbitrary, fixed distribution, and τ is taken in small, discrete timesteps. The argument for this process is as follows: If we add up a large number of independent random variables, each following the same given distribution, the sum will approximate a normal distribution, no matter how the individual random variables were distributed.⁵ This approximate normal distribution is the result of the central limit theorem of probability and why normal distributions are seen so often in nature, from the height of people to the velocity of components of atoms in a gas. If we consider the growth of submission popularity as a large number of random events increasing the logarithm of the popularity by a small, random amount, we arrive at the log-linear model just described.

What follows from the model is that on the natural, linear scale of popularities we must multiply the actual popularity by a small, random amount to obtain the popularity for the next timestep. This process is called “growth with random multiplicative noise,” an unexpected characteristic of the dynamics of user-submitted content.⁹ While the increments at each timestep are random, their expectation value over many timesteps adds up, ultimately to $\ln r(t_0, t_r)$ in the log-linear model. Thus the innate differences among the user-perceived interesting-

ness of submissions should be seen early on, up to a variability accounted for by the noise terms.

Popularity prediction. To illustrate how a content provider might use the random logarithmic growth model of content popularity on Digg and YouTube, we performed straightforward extrapolations on the data we collected to predict future access rates. If submissions do not get more or less attractive over time as they were in the past, we expect their normalized popularity values to follow the trends in Figure 1. The strong correlation between early and later times suggests a submission that is popular at the beginning will also be popular later on. The linearity of popularity accrual with a random additive noise on the logarithmic scale also allows us to approximate the number of views/diggs at any given time in the future; they are predicted to be a constant product of the popularity measured at an earlier time. However, the multiplier depends on when the sampling and the prediction are performed.

In order to perform and validate the predictions, we subdivided the submission time series data into a training set and a test set. For Digg, we took all stories submitted during the first half of the data-collection period (July to mid-September 2007) as the training set and the second half as the test set. On the other hand, the 7,146 YouTube videos we followed were submitted at about the same time, so we randomly selected 50% of them as training and the other 50% as test; the table here outlines the numbers of submissions in the two sets. The linear regression coefficients between t_i and t_r data were determined on the training set, then used to extrapolate on the test set.

Content popularity counts are often related to other quantities, like click-through rates of linked adver-

tisements and number of comments the content is expected to generate on the community site. For this reason we measured the performance of the predictions as the average relative squared error over the test set or as the expected difference of a prediction from the actual popularity, in percentages. For a reference time of t_i to predict the popularity of submissions we chose 30 days after submission time. Since the predictions naturally depend on t_i and how close we are to the reference time, we performed the parameter estimations in hourly intervals starting immediately after the introduction of a submission. The parameter values for the predictions ($\ln r(t_i, t_r)$ in the log-linear model discussed earlier) can be obtained with maximum likelihood fitting from the training-set data.

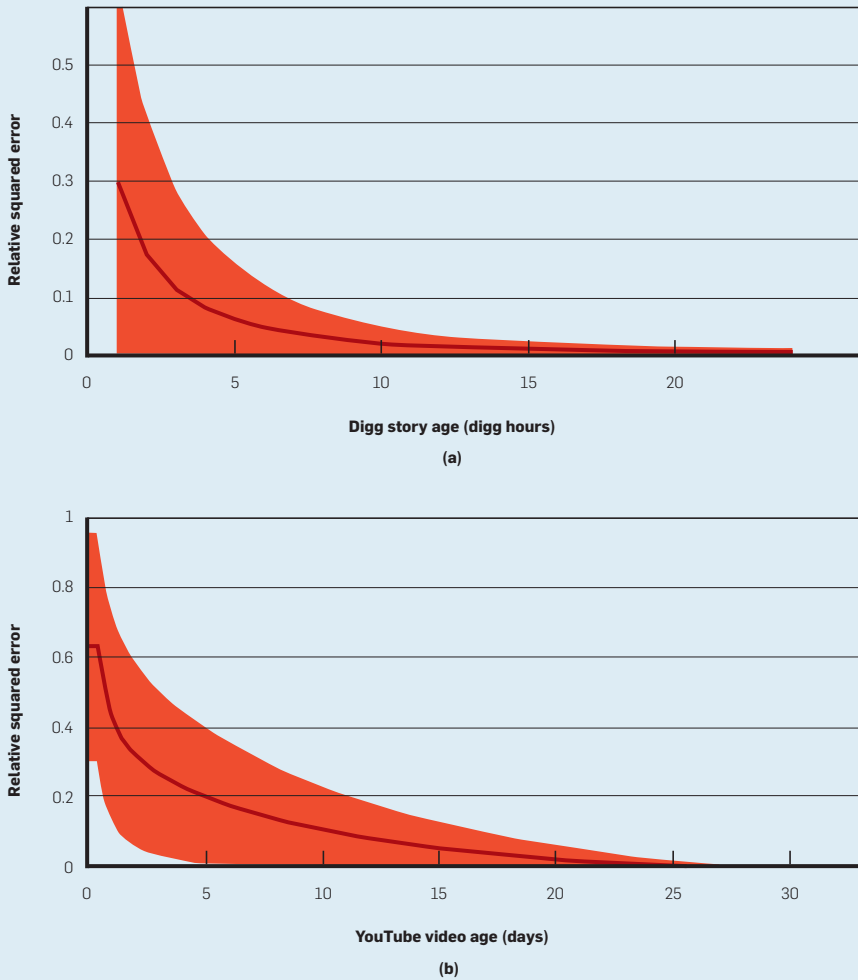
The errors measured on the test set (see Figure 5) show that the expected error decreases rapidly for Digg (negligible after 12 hours), while for YouTube the predictions converge more slowly to the actual value. After five days, the expected error made in estimating the view count of an average video was about 20%, while the same error was attained an hour after a Digg submission. This is due to the fact that Digg stories have a much shorter life cycle than YouTube videos, and Digg submissions quickly collect many votes right after being promoted.

The simple observation that the popularities of individual items are linearly related to each other at different times enables us to extrapolate to future popularities by measuring content popularity shortly after the content is introduced. However, the detailed parameter-estimation procedure strongly depends on the idiosyncrasies of the random multiplicative model and the type of error measure we wish to minimize (such as absolute and relative), so

Partitioning the collected data into training and test sets, we divided the Digg data by time and chose the YouTube videos randomly for each set, respectively.

	Training set	Test set
Digg	10,825 stories (7/1/07–9/18/07)	6,272 stories (9/18/07–12/16/07)
YouTube	3,573 videos randomly selected	3,573 videos randomly selected

Figure 5. Prediction performance is based on the logarithmic growth model measured by the average relative squared error function for (a) Digg and (b) YouTube, respectively. The shaded areas indicate one standard deviation of the individual submission errors around the average.



these constraints must be considered to achieve the minimum error possible allowed by the model.

Social Networking

Social networking features in Web 2.0 services are so ubiquitous it is almost mandatory for a site to offer them to its users. For example, Digg’s approach to social networking is to make it possible for users to be fans of other users, after which they are able to see what stories their “idols” submit or digg. This is essentially a restricted form of collaborative filtering, but users themselves select the peers they wish to follow. A similar kind of social network is active in YouTube, though the feature that allowed users to follow the videos their friends were watching was nascent in

2008; however, they might have seen if friends recently uploaded videos. Due to the limited nature of social-networking options on YouTube in 2008, we focus on the network of Digg users. Together with content-popularity data, we also collected link information using the Digg API. Figure 6 shows a typical snapshot of the Digg social network in 2007, with about 260 users and 550 links, where a link represents whether a particular user is a fan of another user. Users who digg a particular story are in red, with no apparent clustering among them. However, these users are relatively dense in the neighborhood of the small social graph in Figure 6, since the story attracted nearly 15,000 diggs altogether, considerably more than the average submission at the time.

It was known that the Digg social network plays an important role in making a story visible and popular when the submission is still in Digg’s “upcoming” section, with new stories appearing at the top of the “upcoming” page on average every ninth second, as in Figure 2, with about 400 new submissions an hour in 2007. Though all new submissions are shown in the “upcoming” section, the list is updated so quickly that entries left the first page in about two minutes. The most effective way to discover new stories should thus be through the social network, where recent diggs of a user’s idols are visible for more time on the user’s personal page. To what extent then, do diggers pay attention to what their idols already digg?

To see how Digg social networking functioned we took all submissions for which we had data for at least 12 hours after promotion and measured the fraction of diggers with at least one digger among their idols and who had already digg the same story. In essence, this measurement is the probability that a new digg is made by users who may have seen the story through their social networks. We normalized the times of diggs with respect to the promotion time of the individual submissions, so for diggs made before promotion, time is measured backward. Results are outlined in Figure 7, where about 20% of diggers have an idol who digg the same story before they did, when it was still in the “upcoming” phase. However, this figure drops considerably (to 7%) after promotion; most diggs are cast by users who could not have seen the submission in their social network before. This falloff in peer following supports the assumption that stories are found through the social network in the “upcoming” phase, but once they are promoted to the front page and exposed to a diverse audience for a longer time, the effect of the social network becomes negligible.

While users are about three times more likely to digg a submission their idols digg in the “upcoming” phase than after it was promoted, the measurement only intuitively suggests that users pay attention to the activities of their peers. To determine whether diggers are truly influenced by their social peers, the null hypothesis for user

diggs would be a scenario in which users pick stories randomly, never being influenced by what their idols did before them. If the observed fractions substantially exceed the random expectation, we can safely say that users indeed pick the same submissions as their peers.

We were able to test whether users digg stories according to the random null hypothesis by randomly shuffling their activities. We simulated a scenario whereby users made their diggs at exactly the same times as they would in real time to mimic the sessions when they're logged onto Digg. However, we let the agents representing the users digg any story present in the system, rather than what users actually dugg. This approach ensured that the simulated agents picked a random story from among all the stories available to them. We maintained (important) the agents' social links and corresponding user links, so we could observe the presence or lack of a social-network effect. After the agents' selections were randomly made, we performed the same measurement as in Figure 7 to determine how agents might have been influenced by their idols to digg the same stories as their idols.

In Figure 7, the difference between the random model (green line) and the observed digging pattern (blue line) is obvious: Users digging stories in the "upcoming" phase were more than twice as likely to digg what their idols digg than they would if there was no social network—the same as picking a story randomly. However, and most important, stories late in the "promoted" phase get diggs from users who do not watch their links at all; the random hypothesis delivers the same fractions as the real observations after about a day. However, right after promotion, users seem to do the opposite of their peers: The probability that a new digger of a story is a fan of a previous digger of a story is significantly less than one would expect from random choice. The controversy in this result might be resolved if we consider that once a submission is promoted to the front page (shortly after time 0 in the figure), it gains tremendous visibility compared to the "upcoming" phase and is exposed to many casual Digg users. These users do not actively participate in discovering new submis-

sions but browse the Digg main page to see what other users found interesting (making up the bulk of the user base), and, though they do not digg often, their compounded activity dominates the diggs a story gets at this stage. At the same time, they are unlikely to have an

extended (or even any) social network. Consequently, the observed probability of peer influence is diminished.

The beneficial effect of a social network on content popularity is therefore confined to less active periods of the content's life cycle; that is, it matters

Figure 6. Representative example of a Digg-user social network. We randomly selected a user as origin and included every other user in the social graph with snowball sampling up to distance four from the user following breadth-first search. Diggers of a particular story are in red; non-diggers are in green.

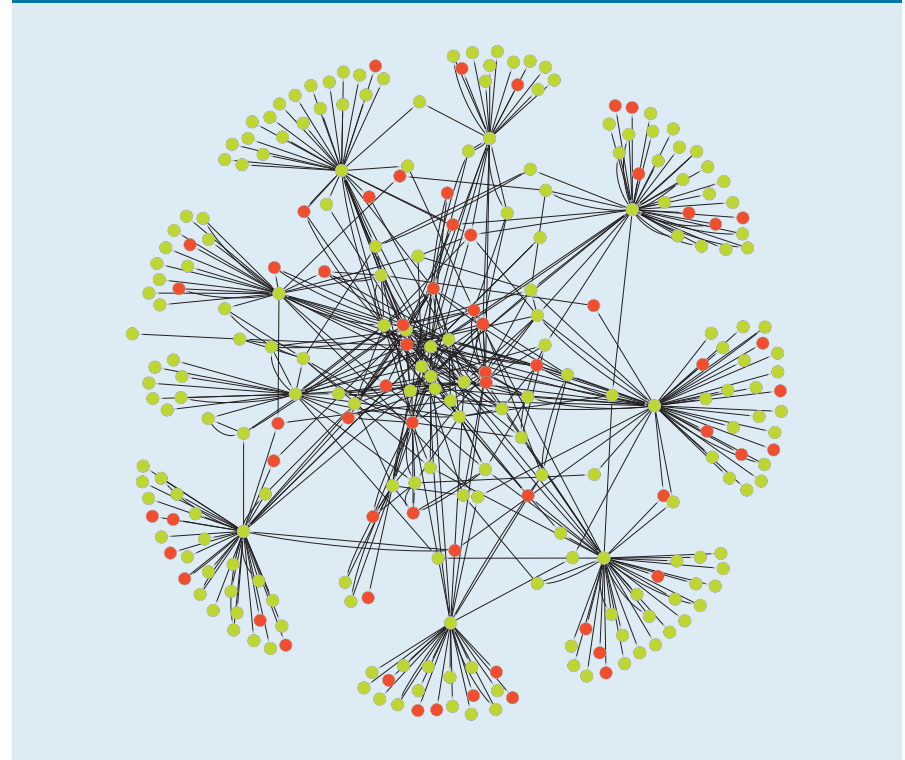
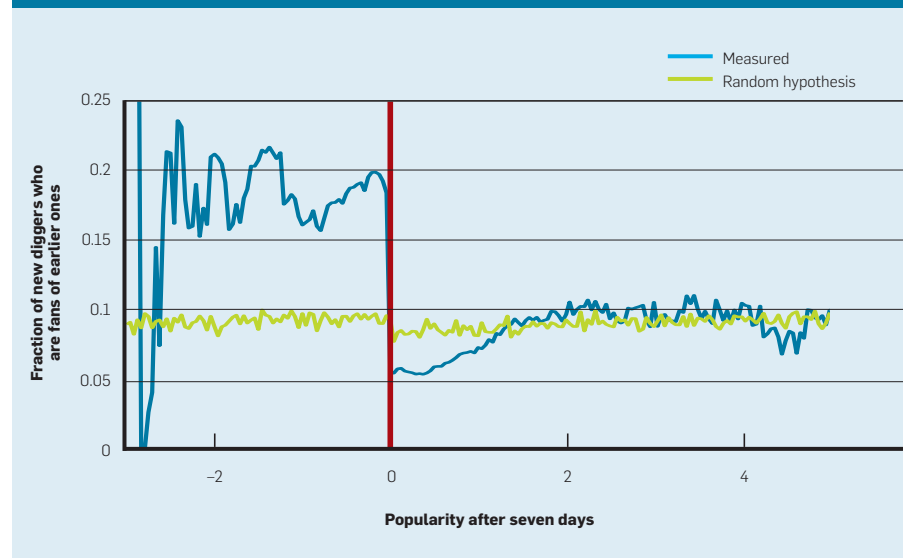


Figure 7. Probability that a digger of a story is a fan of a digger who dugg the same story (blue line) as a function of the time of the digg. Time is relative to the promotion time of the story, with the average calculated over all diggs on all stories. The vertical red line marks time 0 (promotion time), and negative times refer to the "upcoming" phase. The green line is the same measurement but with diggs randomly shuffled.




only when its visibility is minuscule compared to its other stages, and the highest number of diggs accrues when the social-network effect is nonexistent. We therefore do not consider this feature (otherwise deemed important) a main contributor from a prediction point of view in terms of total popularity count.

Conclusion


In this article we have presented our method for predicting the long-term popularity of online content based on early measurements of user access. Using two very popular content-sharing portals—Digg and YouTube—we showed that by modeling the accrual of votes on and views of content offered by these services we are able to predict the dynamics of individual submissions from initial data. In Digg, measuring access to given stories during the first two hours after posting allowed us to forecast their popularity 30 days ahead with a remarkable relative error of 10%, while downloads of YouTube videos had to be followed for 10 days to achieve the same relative error. The differing time scales of the predictions are due to differences in how content is consumed on the two portals; Digg stories quickly become outdated, while YouTube videos are still found long after they are submitted to the portal. Predictions are therefore more accurate for submissions for which attention fades quickly, whereas predictions for content with a longer life cycle are prone to larger statistical error.

We performed experiments showing that once content is exposed to a wide audience, the social network provided by the service does not affect which users will tend to look at the content, and social networks are thus not effective promoting downloads on a large scale. However, they are important in the stages when content exposure is constrained to a small number of users.

On a technical level, a strong linear correlation exists between the logarithmically transformed popularity of content at early and later times, with the residual noise on this transformed scale being normally distributed. Based on our understanding of this correlation, we presented a model to be used to predict future popularity, comparing its performance to the data we collected.



In the presence of a large user base, predictions can be based on observed early time series, while semantic analysis of content is more useful when no early click-through information is available.



We thus based our predictions of future popularity only on values measurable at the time we did the study and did not consider the semantics of popularity and why some submissions become more popular than others; however, this semantics of popularity may be used to predict click-through rates in the absence of early-access data.⁸ In the presence of a large user base, predictions can be based on observed early time series, while semantic analysis of content is more useful when no early click-through information is available.

However, we could not explore several related areas here. For example, it would be interesting to extend the analysis by focusing on different sections of the portals (such as how the YouTube “news & politics” section differs from the YouTube “entertainment” section). We would also like to learn whether it is possible to forecast a Digg submission’s popularity when the diggs come from only a small number of users whose voting history is known, as it is for stories in Digg’s “upcoming” section. **□**

References

1. Alexa Web Information Service; <http://www.alexa.com>
2. Cha, M., Kwak, H., Rodriguez, P., Ahn, Y.-Y., and Moon, S. I tube, you tube, everybody tubes: Analyzing the world’s largest user-generated content video system. In *Proceedings of the Seventh ACM SIGCOMM Conference on Internet Measurement* (San Diego, Oct. 24–26). ACM Press, New York, 2007, 1–14.
3. Cheng, X., Dale, C., and Liu, J. Statistics and social network of YouTube videos. In *Proceedings of the 16th International Workshop on Quality of Service* (Enschede, The Netherlands, June 2–4, 2008), 229–238.
4. Digg API; <http://digg.com/api/docs/overview>
5. Feller, W. *An Introduction to Probability Theory and Its Applications, Vol. I*. John Wiley & Sons, Inc., New York, 1968.
6. Gill, P., Arlitt, M., Li, Z., and Mahanti, A. YouTube traffic characterization: A view from the edge. In *Proceedings of the Seventh ACM SIGCOMM Conference on Internet Measurement* (San Diego, Oct. 24–26). ACM Press, New York, 2007, 15–28.
7. Lerman, K. Social information processing in news aggregation. *IEEE Internet Computing (Special Issue on Social Search)* 11, 6 (Nov. 2007), 16–28.
8. Richardson, M., Dominowska, E., and Ragno, R. Predicting clicks: Estimating the click-through rate for new ads. In *Proceedings of the 16th International Conference on the World Wide Web* (Banff, Alberta, Canada, May 8–12). ACM Press, New York, 2007, 521–530.
9. Wu, F. and Huberman, B.A. Novelty and collective attention. *Proceedings of the National Academy of Sciences* 104, 45 (Nov. 2007).
10. YouTube API; <http://code.google.com/apis/youtube/overview.html>

Gabor Szabo (gabors@hp.com) is a research scientist in the Social Computing Lab at Hewlett-Packard Labs, Palo Alto, CA.

Bernardo A. Huberman (bernardo.huberman@hp.com) is an HP Senior Fellow and Director of the Social Computing Lab at Hewlett-Packard Labs, Palo Alto, CA.