Social Information Processing in News Aggregation

Social media sites underscore the Web's transformation to a participatory medium in which users collaboratively create, evaluate, and distribute information. Innovations in social media have led to social information processing, a new paradigm for interacting with data. The social news aggregator Digg exploits social information processing for document recommendation and rating. Additionally, via mathematical modeling, it's possible to describe how collaborative document rating emerges from the independent decisions users make. Using such a model, the author reproduces observed ratings that actual stories on Digg have received.

Kristina Lerman
University of Southern California

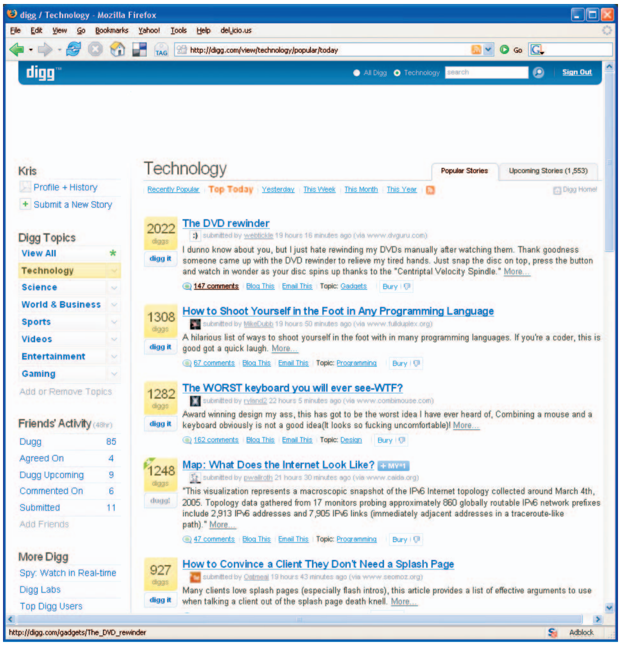
The label *social media* has been attached to a growing number of Web sites whose content is primarily user driven, including blogs, wikis, and media sharing sites such as Flickr, del.icio.us, and Digg (which let users share, discuss, and rank photos, Web pages, and news stories, respectively). Other sites, such as Amazon's Mechanical Turk, let users collaboratively find innovative solutions to hard problems. Social media's rise underscores a transformation in the Web that's as fundamental as its birth: rather than simply searching for and passively consuming information, users are collaboratively creating, evaluating, and distributing it. In the near future, social media could enable new information-processing applications that will allow personalized information discovery, exploit the "wisdom of crowds" (such as emergent semantics and collaborative information evaluation), help us more deeply analyze community structure to identify trends and experts, and provide further functions still difficult to imagine.

Social media sites share four characteristics:

- users create or contribute content in various media types;
- users annotate content with tags;
- users evaluate content, either actively by voting or passively by using it; and
- users create social networks by desig-

1089-7801/07/$25.00 © 2007 IEEE

nating other users with similar interests as contacts or friends.



2022 diggs digg it

The DVD rewinder

submitted by webtickle 19 hours 16 minutes ago (via www.dvguru.com)

I dunno know about you, but I just hate rewinding my DVDs manually after watching them. Thank goodness someone came up with the DVD rewinder to relieve my tired hands. Just snap the disc on top, press the button and watch in wonder as your disc spins up thanks to the "Centriptal Velocity Spindle." More...

147 comments | Blog This | Email This | Topic: Gadgets | Bury

1308 diggs digg it

How to Shoot Yourself in the Foot in Any Programming Language

submitted by MikeDubb 19 hours 50 minutes ago (via www.fullduplex.org)

A hilarious list of ways to shoot yourself in the foot with in many programming languages. If you're a coder, this is good got a quick laugh. More...

67 comments | Blog This | Email This | Topic: Programming | Bury

1282 diggs digg it

The WORST keyboard you will ever see-WTF?

submitted by ryland2 22 hours 5 minutes ago (via www.combimouse.com)

Award winning design my ass, this has got to be the worst idea I have ever heard of, Combining a mouse and a keyboard obviously is not a good idea(It looks so fucking uncomfortable)! More...

162 comments | Blog This | Email This | Topic: Design | Bury

1248 diggs dugg!

Map: What Does the Internet Look Like? + MY#1

submitted by pwallroth 21 hours 30 minutes ago (via www.caida.org)

"This visualization represents a macroscopic snapshot of the IPv6 Internet topology collected around March 4th, 2005. Topology data gathered from 17 monitors probing approximately 860 globally routable IPv6 network prefixes include 2,913 IPv6 addresses and 7,905 IPv6 links (immediately adjacent addresses in a traceroute-like path)." More...

47 comments | Blog This | Email This | Topic: Programming | Bury

927 diggs digg it

How to Convince a Client They Don't Need a Splash Page

submitted by Oatmeal 19 hours 43 minutes ago (via www.seomoz.org)

Many clients love splash pages (especially flash intros), this article provides a list of effective arguments to use when talking a client out of the splash page death knell. More...

Figure 1. Typical Digg page. Clicking on a story's title takes users to the source, whereas clicking on "diggs" takes them to a page describing the story's activity.

Through these characteristics, social media facilitate new ways of interacting with information and enhance collaborative problem solving through what I call *social information processing*. To demonstrate this concept, I discuss how the social news aggregator Digg (www.digg.com) uses social information processing to solve two long-standing information-processing problems: document recommendation and rating. I present a mathematical model that describes the dynamics of collaborative rating of news stories. The model takes into account the influence users exert through their social networks, and it correctly predicts the observed behavior of ratings that actual stories on Digg have received.

Recommendation and Rating

Social filtering or *social recommendation* is an effective alternative to collaborative filtering (CF), a popular recommendation technology that commercial giants such as Amazon and Netflix use. CF-based recommender systems ask users to express their opinions by rating items, and then suggest new items that users with similar opinions also liked. One noted problem with CF is that users are generally resistant to rating.[1] In contrast, users express their tastes and preferences on social media sites by creating personal networks comprising tens to hundreds (even thousands) of friends. Researchers have studied social recommendation in the context of innovation and viral marketing,[2,3] helping advertisers target their messages[4] to get the most from the "word of mouth" effect. This article investigates the inverse problem — how people can use social networks as a recommendation mechanism to effectively filter vast streams of information.[5,6]

Another outstanding problem in information processing is how to evaluate a document's quality. This issue crops up daily in information retrieval and Web search, during which users aim to find — among the terabytes of data accessible online — information most relevant to their queries. Search engines' standard practice is to identify all documents using terms that appear in the user's query and rank the results according to their quality or importance. Google revolutionized Web search by exploiting the Web's link structure — created via Web page authors' independent activities — to evaluate Web page content.[7] Similarly, social news aggregators such as Digg and Reddit (http://reddit.com) rely on their users' opinions to evaluate quality of news stories.

Digg Anatomy

Digg's functionality is very simple: users submit stories they find online, and other users rate these stories by voting. Digg also lets users create personal social networks by adding other users as friends and provides an interface to easily track those friends' activities, such as what stories they read and liked. Each day, Digg promotes a handful of stories to its front page. Rather than depend on a few editors' decisions, the promotion mechanism emerges from the voting patterns of many users. This type of collective decision-making can be extremely effective in breaking news, often outperforming special-purpose algorithms. In a talk presented at the Web 2.0 conference in November 2006, Digg's founder Kevin Rose reported that the news of Donald Rumsfeld's resignation in the wake of the 2006 US Congressional elections broke Digg's front page within three minutes of submission and 20 minutes before Google News relayed it. In addition to selecting popular news stories, Digg ranks users by how successful they are at getting their stories promoted to the front page.

Figure 1 shows a typical Digg page. When a user submits a story, it goes into the upcoming stories queue. One to two new submissions come in every minute, and they're displayed in reverse chronological order (most recent at the top), 15 stories to a page. The story's title is a link to the source, whereas clicking on the number of diggs (votes) the story received takes you to a page describing the story's Digg activity (the discussion around it, the list of people who voted on it, and so on).

To vote on a story, a user "diggs" it (clicks the "digg it" button), which also saves it to the user's history. Digg also lets users "bury" stories that are spam or duplicates, or that contain inappropriate materials. Burying a story doesn't reduce its rating the way voting a story down does on Reddit; rather, if enough people have buried a story, Digg removes it permanently.

Emergent Front Page

When a story gets enough votes, Digg promotes it to the front page. Most people who visit Digg daily, or subscribe to its RSS feeds, read only the front-page stories; thus, getting to the front page greatly increases a story's visibility. Although the exact promotion mechanism is a secret and changes periodically, it appears to take into account how many votes the story receives. Digg's popularity is fueled in large part by this *emergent front page* phenomenon.

Other social media sites rely on similar mechanisms to showcase select content. Every day, the photo-sharing site Flickr (www.flickr.com) chooses the 500 most "interesting" of the newly uploaded images to feature on its Explore page. The selection algorithm takes into account how many times users viewed the image, commented on it, or marked it as a favorite (www.flickr.com/explore/interesting/). Therefore, Flickr's Explore page also arises from decisions many users have made. Similarly, the social bookmarking site del.icio.us (http://del.icio.us) showcases the most popular of the recently tagged Web pages.

Social Networking

Digg lets users designate others as friends and makes it easy to track their activities. The Friends interface on the page's left column summarizes how many stories a user's friends have submitted, commented on, or dugg recently. Tracking friends' activities is a common feature of many social media sites and is one of their major draws. It also offers a new paradigm for interacting with information: rather than actively searching for new interesting content or subscribing to a set of predefined topics, users can let others find and filter information for them – what I call social filtering.

Top Users

Until February 2007, Digg ranked users according to how many of their stories it promoted to the front page – that is, the user who was ranked number one had submitted the most front-page stories, the number two user had fewer stories promoted to the front page, and so on. Clicking on the Top Users link displays the ranked list of users. Speculation exists that ranking users increases competition, leading some to be more active in order to improve their rankings. Digg discontinued making the Top Users list publicly available, citing concerns that marketers were paying these users to promote their products and services.[8]

Social Filtering

In order to study social networks' role in information recommendation and filtering, my team at USC tracked both upcoming and front-page stories in Digg's technology section by scraping the Digg site with help from Web wrappers we created using Fetch Technologies tools (www.fetch.com):

- The *digg-frontpage* wrapper extracts a list of the 210 most recently promoted stories. For each story, it extracts the submitter's name, the story title, when it was submitted, and how many votes and comments it received, along with the names of the first 216 users who voted on it.
- The *digg-all* wrapper extracts a list of the 300 most recently submitted stories. For each story, it extracts the submitter's name, the story title, the time it was submitted, and how many votes and comments it received.
- The *top-users* wrapper extracts information about the top 1,020 recently active users. For each user, it extracts how many stories that user has submitted, commented, and voted on; how many of that user's submitted stories have been promoted to the front page; the user's rank; and his or her friends list and reverse friends – that is, "people who have befriended this user."

We executed digg-frontpage and digg-all wrappers hourly over a period of several days in May and June 2006. We executed the top-users wrapper weekly starting in July 2006 to gather snapshots of the top Digg users' social networks.

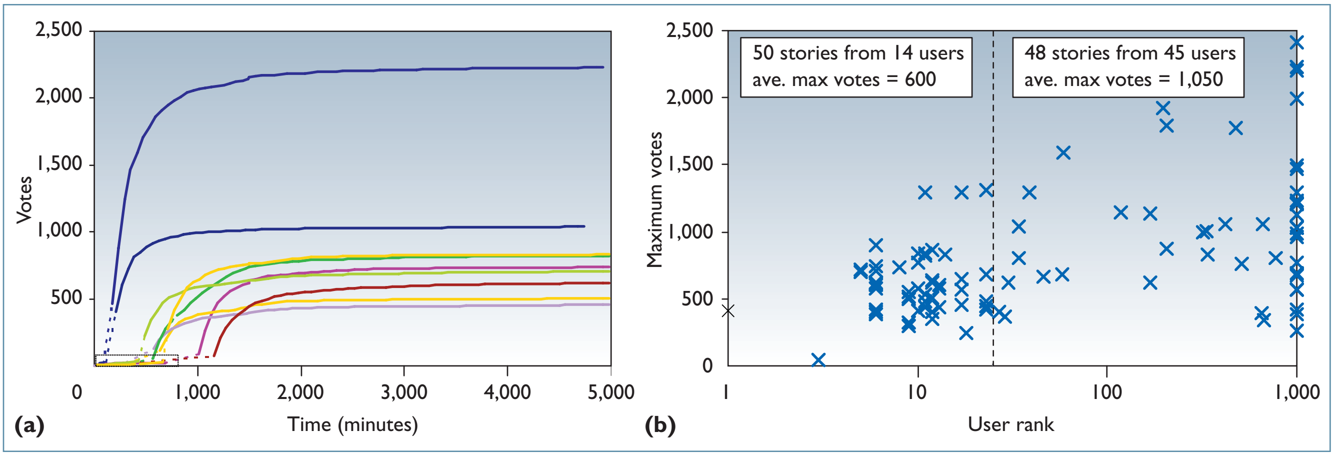


Figure 2. Vote dynamics and distribution. (a) The number of votes received by select stories over four days. The small rectangle indicates the period the stories were in the upcoming stories queue, whereas dashes indicate the story's transition to the front page. (b) We also recorded the maximum votes stories received during the period of observation vs. the submitter's rank. The symbols on the right axis correspond to low-rated users with rank > 1,020.

We identified stories users submitted to Digg throughout roughly one day in May 2006 and followed them over a period of several days. Of the 2,858 stories that 1,570 distinct users submitted during this time period, only 98 stories from 60 users made it to the front page. Figure 2a shows select stories' ratings evolution (number of votes). All stories' basic dynamics appear the same. While in the upcoming queue, a story accrues votes at a slow rate. Once Digg promotes the story to the front page, it accumulates votes much more quickly. As the story ages, its accumulation of new votes slows,[9] and the story's rating saturates at some value, which we call *interestingness*, that indicates how interesting the story is to the general Digg community.

Note that the top-ranked users aren't necessarily submitting the stories that get the most votes. Figure 2b displays the maximum votes a story has received versus the story submitter's rank. Slightly more than half the stories came from 14 top-ranked users (rank < 25), whereas 48 came from 45 low-ranked users. Stories the top-ranked users submitted have an average interestingness of 600, almost half the average interestingness of stories the low-ranked users submitted. Note that top-ranked users are responsible for multiple front-page stories. A look at the Top Users list shows that this is generally the case: of the more than 15,000 front-page stories submitted by the top 1,020 users as of May 2006, the top 3 percent of the users were responsible for 35 percent of the stories.

Social Networks and Recommendation

If top-ranked users aren't submitting the most interesting stories, why are they so successful? I believe that social filtering plays a role in helping promote stories to the front page. As I explained earlier, Digg lets users track friends' activities, including the stories they've submitted, commented, and voted on. I claim that users employ the Friends interface to filter the tremendous number of new submissions on Digg to find new interesting stories.

The friend relationship is asymmetric: when user A lists user B as a friend, A can watch B's activities, but not vice versa. A is thus B's reverse friend. Figure 3a shows a scatter plot representing the top 1,020 Digg users' friends versus their reverse friends as of May 2006. The orange symbols correspond to the top 33 users. For the most part, users appear to exploit Digg's social networking feature, with the top users having bigger social networks. Two of the biggest celebrities (those that the most people watch) are users *a* and *b* in Figure 3a. These users correspond to *kevinrose* and *diggnation*, respectively — one of Digg's founders and a user account set up for the popular Digg story podcast.

First, I present indirect evidence of social filtering by showing that users' success rates correlate to their social network size. I define a user's success rate as the fraction of stories the user submitted that Digg promoted to the front page. I use statistics about the top 1,020 users' activities to show that those with bigger social networks are more successful at getting their stories promoted. In the analysis, I include only users who have submitted 50 or more stories (514 users). Figure 3b shows the correlation between users' mean success rates and the size of their social networks. (I binned

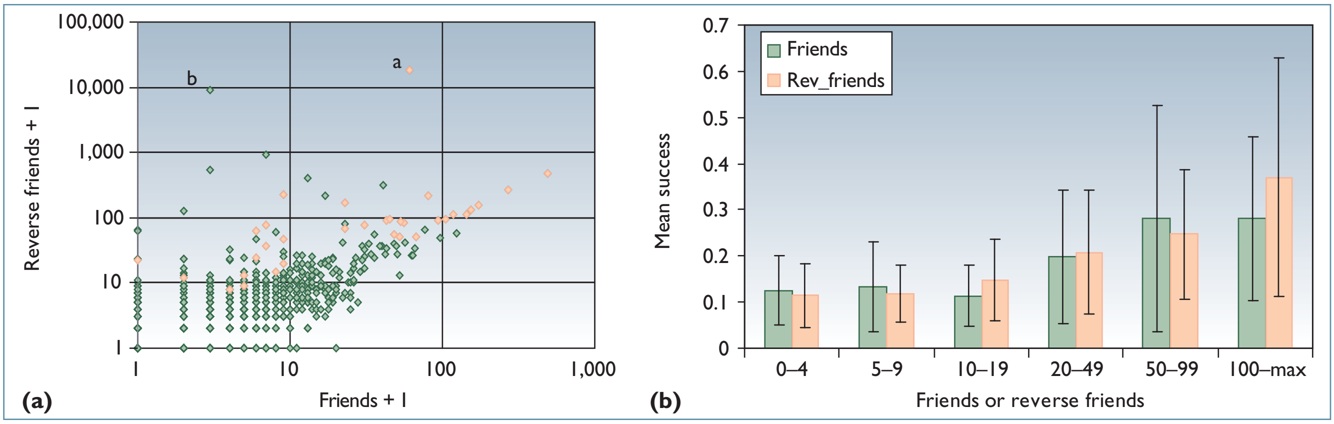


Figure 3. Usage statistics. (a) A scatter plot represents the top 1,020 Digg users' friends vs. their reverse friends. (b) We see the strength of the linear correlation coefficient between a user's success rate and how many friends and reverse friends he or she has.

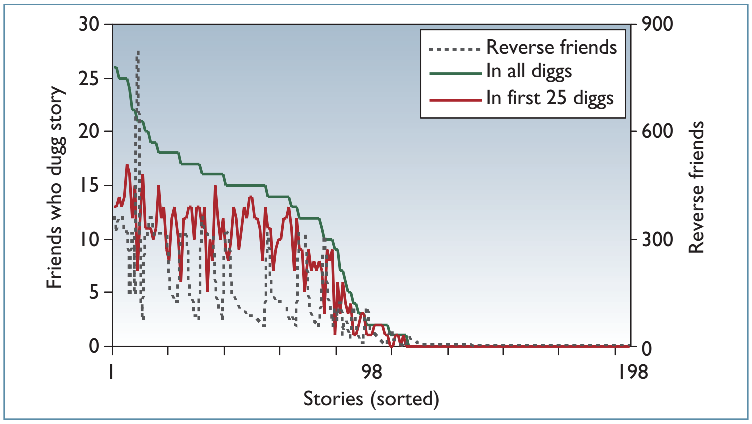


Figure 4. Users who liked stories their friends submitted. The dashed line shows the size of a story submitter's social network, whereas the green line shows how many of the first 216 votes came from his or her social network, and the red line shows how many of the first 25 votes came from within the network.

data to improve statistics.) Despite large error bars, a significant correlation exists between users' success rates and the size of their social networks — and more importantly, the number of reverse friends they have.

Next, I present additional evidence that users utilize the Friends interface to find new interesting stories. I analyze two subclaims: that users digg stories their friends submit and that they digg stories their friends digg.

Users digg stories their friends submit. To show that users digg stories their friends submit, I used the digg-frontpage wrapper to collect 195 front-page stories, each listing the first 216 users to vote on the story (15,742 distinct users total); the submitter's name is first on the list. We can compare this list, or any portion of it, with the list of the submitter's reverse friends. The dashed line in Figure 4 shows the size of the submitter's social network (number of reverse friends). More than half the stories (99) were submitted by users with more than 20 reverse friends and the rest by poorly connected users. (These users have rank > 1,020 and weren't listed as friends of any of the 1,020 users in our data set; it's possible, though unlikely, that they have reverse friends.) The green line shows how many voters are also among the submitter's reverse friends. All but two of the stories submitted by well-connected users received diggs from the submitter's reverse friends.

We can use simple combinatorics[10] to compute the probability that k of the submitter's reverse friends could have voted on the story purely by chance. The probability that, after picking $n = 215$ users randomly from a pool of $N = 15{,}742$, you end up with k from a group of size K is $P(k, n) = \binom{n}{k}(p)^k (1 - p)^{n-k}$, where $p = K/N$. Using this formula, the probability (averaged over stories that at least one friend dugg) that the observed numbers of reverse friends voted on the story by chance is $P = 0.005$. (If we include in the average the two stories that none of the submitter's friends dugg, we end up with a higher but still significant $P = 0.023$.) Moreover, users digg stories that their friends submit very quickly. The red line in Figure 4 shows the number of reverse friends who were among the first 25 voters. The probability that we could have observed these numbers by chance is even less: $P = 0.003$. We conclude that users digg the stories their friends submit. As a side effect, by letting users quickly digg stories that friends sub-

Table 1. Stories submitted by poorly connected users.

Diggers	*m* = 1	*m* = 6	*m* = 16	*m* = 26	*m* = 36	*m* = 46
Visible to friends*	34	75	94	96	96	96
Dugg by friends**	10	23	37	46	49	55
Probability***	0.005	0.028	0.060	0.077	0.090	0.094

* through the Friends interface

** of the first *m* diggers within the next 25 diggs

*** that the friends dugg the stories by chance

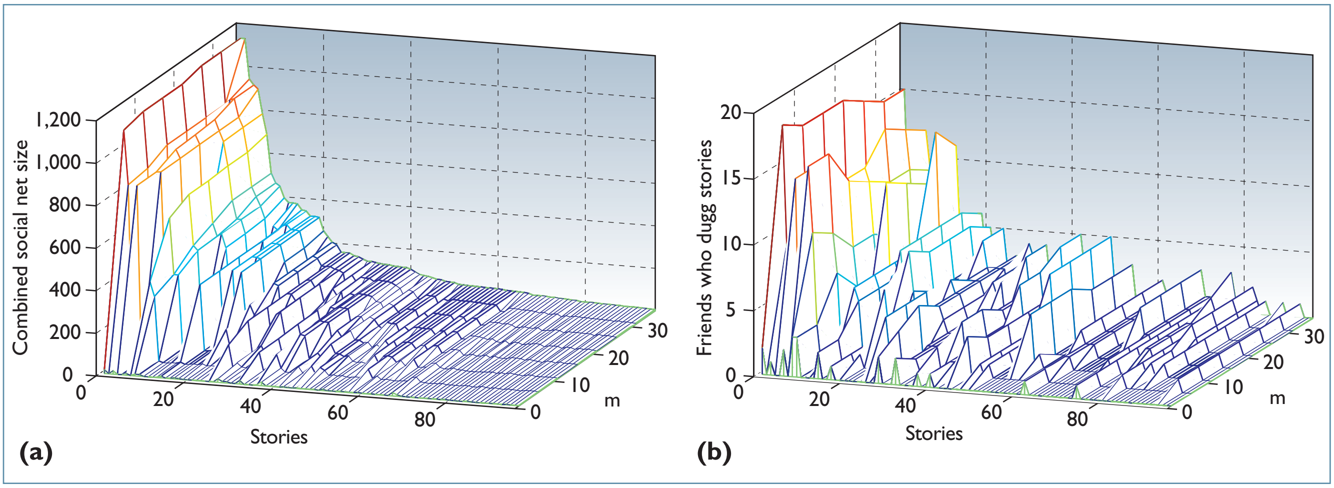


Figure 5. Stories from poorly connected users. The plots show (a) the combined social network size of the first m *users who voted on the story, and (b) the number of the 25 subsequent votes that came from the first* m *voters' friends.*

mit, social networks play an important role in promoting stories to the front page.

Users digg stories their friends digg. Do social networks also help users discover interesting stories that poorly connected users submit? Digg's Friends interface lets users see stories their friends have dugg. As a well-connected user diggs stories that users with few or no reverse friends have submitted, are others within his or her social network more likely to read them?

Figure 5 shows how well-connected users' activities affected the 96 stories submitted by poorly connected users, or those with fewer than 20 reverse friends; *m* = 1 corresponds to the story submitter, whereas *m* = 6 corresponds to the story's submitter and the first five users to digg it. Figure 5a shows how the first *m* diggers' combined social network (number of reverse friends) grows as a story receives votes. Figure 5b shows how many of the subsequent 25 votes came from users within the first *m* voters' combined social network.

At the time of submission (*m* = 1), only 34 of the 96 stories were visible to others within the submitter's social network, and the submitter's reverse friends dugg 10 of these within the first 25 votes. After 15 more users had voted, almost all the stories were visible through the Friends interface. Table 1 summarizes these observations and gives the probability that the observed numbers of reverse friends voted on the story purely by chance. The probabilities for *m* = 26 through *m* = 36 are greater than the 0.05 significance level, possibly reflecting a story's increased visibility on the front page. Although the effect isn't quite as dramatic as that from the previous section, I believe that the data indicates that users do use the "see the stories my friends have dugg" portion of the Friends interface to find new interesting stories.

Changing the Promotion Algorithm

Digg's goal is to feature only the most interesting of the submitted stories on its front page, and it aggregates the opinions of thousands of users —

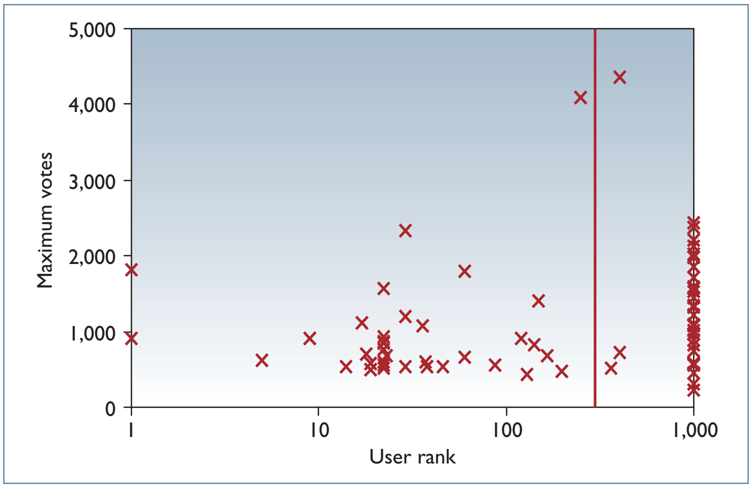


Figure 6. Maximum votes received by front-page stories vs. submitter's rank. My team at USC collected data from stories submitted to Digg in early November 2006, after Digg changed its promotion algorithm. The vertical line divides the set in half. Symbols on the right-hand axis correspond to low-rated users with rank > 1,020.

rather than those of a few dedicated editors – to achieve this goal. I demonstrated in the previous section that social networks play an important role in social filtering and recommendation. Because some users are more active than others, directly implementing social filtering might lead to a "tyranny of the minority" in which the majority of front-page stories came from users with the biggest and most active social networks. A similar finding in September 2006 (see http://taylorhayward.org/digggaming.html) led some Digg users to accuse a "cabal" of top users of gaming the system by automatically voting on each other's stories. The resulting uproar[11] prompted Digg to change the algorithm it uses to promote stories. To discourage what was seen as "bloc voting," the new algorithm looked "at the unique digging diversity of the individuals digging the story" (see http://diggtheblog.blogspot.com/2006/09/digg-friends.htm).

Analysis of the votes that stories submitted in early November 2006 received indicates that the algorithm change did reduce the top-user dominance on the front page. It shows that of the 3,072 stories submitted by 1,865 users over roughly one day, Digg promoted 71 stories from 63 users to the front page. Figure 6 shows the maximum number of diggs these stories received over a six-day period versus the submitter's rank. Compared to the May data (Figure 2b), the front page now demonstrates greater diversity among users, with fewer users responsible for multiple submissions (1.2 stories per submitter compared to 1.6 stories per submitter). Rank distribution is less skewed toward top-ranked users than previously: half the stories came from users with rank < 300, rather than rank < 25, as with the May data set. A smaller spread also exists in the mean interestingness of stories submitted by top- and low-ranked users (960 versus 1,270 in November; 600 versus 1,050 in May). Note that the overall increase in the maximum number of votes stories received could reflect growth in the Digg user base.

Although we can look at these changes as a positive development, they might have unintended consequences – for example, users might not join social networks if they think their votes will be discounted. Rather than being a liability, however, social networks are a useful feature of social media sites, given that individuals can use them to personalize and tailor information.[12] In the next section, I use mathematical analysis to study the behavior of collaborative rating algorithms. Such analysis can help investigate the consequences of changing the promotion algorithm before Digg implements it.

Mathematical Model for Collaborative Rating

We parameterize a story by its interestingness coefficient r, which gives the probability that a story will receive a positive vote once seen. How many votes a story receives depends on its *visibility*, which simply means how many people can see and follow the link to it. A story can receive visibility through the front page, the upcoming stories queue, or the Friends interface.

Visibility on Digg's Pages

A story's visibility on the front page decreases as newly promoted stories push it farther down the list. Although Digg doesn't provide data about its visitors' behavior – specifically, how many proceed to page 2, 3, and so on – I propose to describe it with a simple model that holds that some fraction c_f of visitors to the current front page proceed to the next front page. Thus, if N users visit Digg's front page within some time interval, $c_f N$ users see the second-page stories, and $c_f^{p-1} N$ users see page p stories. A similar model describes how a story's visibility in the upcoming stories queue decreases as new submissions push it down the list. If a fraction c of Digg visitors proceeds to the upcoming stories section and, of these, a fraction c_u proceeds to the next upcom-

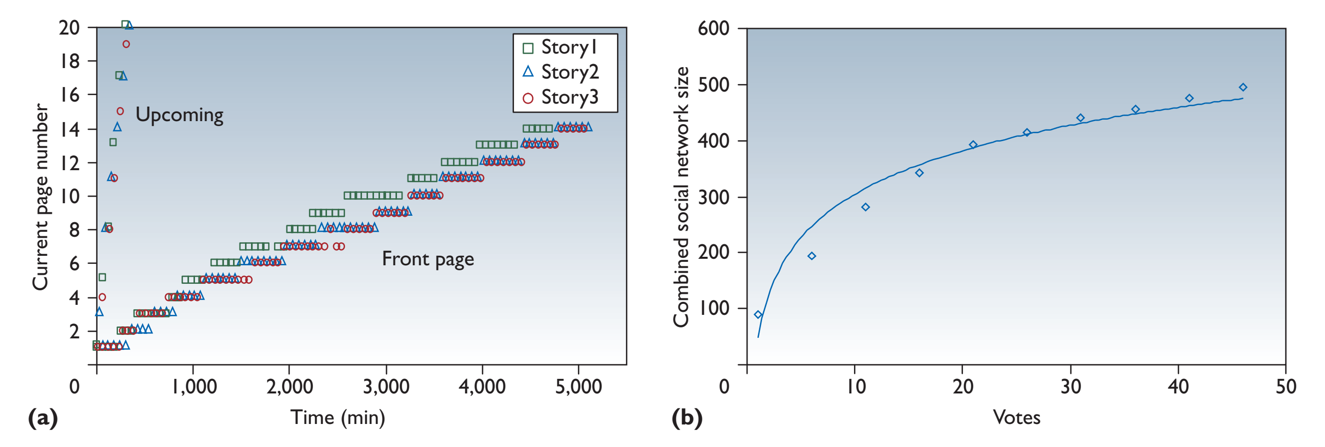


Figure 7. Parameter estimation from data. (a) The current page number of three typical stories in the upcoming stories queue and on the front page vs. time. (b) The average size of S_m*, the combined social network of the first* m *users to digg the story. Although* S_m *is highly variable from story to story, its average value grows consistently:* $S_m = 112.0 * \log(m) + 47.0$.

ing page, then cc_uN of Digg visitors see second-page stories, and $cc_u^{q-1}N$ users see page q stories.

Figure 7a shows how the current page number (in both the upcoming stories queue and the front page) changes over time for three randomly chosen stories from the May data set. The change in a story's current page number is fit well by lines $q, p = k_{u,f}t$ with slopes $k_u = 0.060$ pages per minute (3.60 pages per hour) on the upcoming stories and $k_f = 0.003$ pages per minute (0.18 pages per hour) on the front page.

We use a simple threshold to model how Digg promotes a story to the front page. When the number of votes a story receives is fewer than h, the story is visible on the upcoming pages; when it's greater than h, it's visible on the front pages. This seems to approximate Digg's promotion algorithm as of May 2006, given that in our data set, we didn't see any front-page stories with fewer than 44 votes, nor did we see any upcoming stories with more than 42 votes.

Visibility through the Friends Interface

As mentioned previously, the Friends interface lets users see the stories their friends have submitted, dugg, or commented on during the preceding 48 hours, or see friends' stories that are still in the upcoming stories queue. Although users are probably taking advantage of all four features, I consider only the first two in the analysis. These closely approximate the functionality that other social media sites offer (for example, Flickr lets users see the latest images their friends uploaded, as well as the images friends liked (marked as favorite). I believe that these features are more familiar to users, and that users utilize them more frequently than other ones.

Submitter's friends. Let S be the number of reverse friends a submitter has. As a reminder, these are users who are watching the submitter's activity. We assume that they visit Digg daily, and because they're likely to be geographically distributed across many time zones, they see the submitted story at an hourly rate of $a = S/24$. The story's visibility through the submitter's social network is therefore $v_{\underline{s}} = a\Theta(S - at)\Theta(48 - t)$; $\Theta(x)$ is a step function whose value is 1 when $x \geq 0$ and 0 when $x < 0$. The first step function accounts for the fact that the pool of reverse friends is finite. As users from this pool read the story, the number of potential readers gets smaller. The second step function accounts for the fact that the story will be visible through the Friends interface for 48 hours after submission only.

Voters' friends. As users digg the story, it becomes visible to more users through the "see the stories my friends dugg" part of the Friends interface. Figure 7b shows the average size of S_m, the combined social network of the first m users to digg the story. Although S_m is highly variable from story to story, its average value grows consistently: $S_m = 112.0 * \log(m) + 47.0$. Thus, the story's visibility through the combined social network of the first m users who vote on it is $v_m = bS_m\Theta(h - m)\Theta(48 \text{ hrs} - t)$, where b is a scaling factor that depends on the time interval: for hourly counts, it's $b = 1/24$.

Dynamical Model

In summary, the four factors that contribute to a story's visibility are:

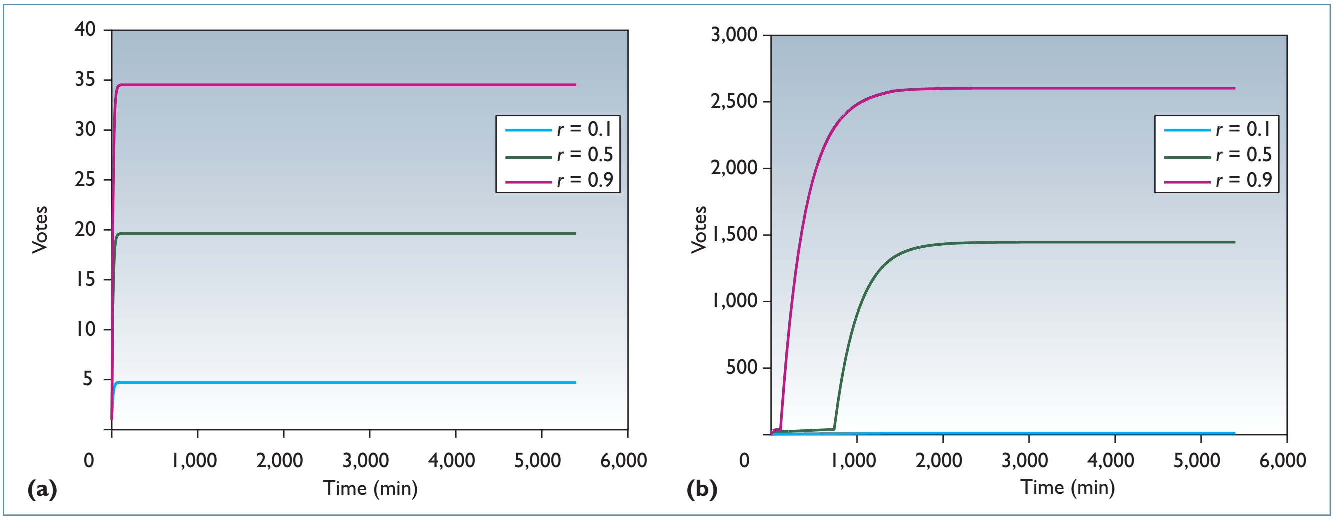


Figure 8. Effect of the submitter's social network on a story's ratings evolution. I examine votes received for a story posted by (a) an unknown user with S = 0 and (b) a connected user with S = 80.

$$v_f = c_f^{p(t)-1} N\Theta\big(m(t)-h\big) \tag{1}$$

$$v_f = cc_u^{q(t)-1} N\Theta\big(h-m(t)\big)\Theta(24hrs-t) \tag{2}$$

$$v_s = a\Theta(S-at)\Theta(48hrs-t) \tag{3}$$

$$v_m = bS_m\Theta\big(h-m(t)\big)\Theta(48hrs-t), \tag{4}$$

where t is time since the story's submission. The first step function in v_f and v_u indicates that when a story has fewer votes than required for promotion, it's visible in the upcoming stories pages; when $m(t) > h$, the story is visible on the front page. The second step function in the v_u term accounts for the fact that a story stays in the upcoming queue for 24 hours, whereas step functions in v_s and v_m model the fact that it's visible in the Friends interface for 48 hours. The story's current page number on the upcoming page q and the front page p change in time according to

$$p(t) = (k_f(t - T_h) + 1)\Theta(T_h - t) \tag{5}$$

$$q(t) = k_u t + 1, \tag{6}$$

with k_u = 0.060 pages per minute and k_f = 0.003 pages per minute. T_h is the time at which the story gets promoted to the front page.

The change in the number of votes m a story receives during a time interval Δt is

$$\Delta m(t) = r\,(v_f + v_u + v_s + v_m)\Delta t. \tag{7}$$

We can solve this equation subject to initial conditions $m(t = 0) = 1$, $q(t = 0) = 1$ because a newly submitted story appears at the top of the upcoming stories queue, and it starts with a single vote coming from the submitter. The initial condition for the front page is $p(t < T_h) = 0$, where T_h is the time at which the story was promoted to the front page. We take Δt to be one minute. Equation 7's solutions show how the number of votes a story received changes in time for different values of parameters c, c_u, c_f, r, and S. Of these, only the last two parameters — the story's interestingness r and the submitter's social network size S — change from one submission to another. Thus, we fix values of the first three parameters $c = 0.3$, $c_u = 0.3$, and $c_f = 0.3$ and study the effect that r and S have on the evolution of the number of diggs a story receives. We also fix the rate at which visitors visit Digg at $N = 10$ users per minute. The actual visiting rate might be vastly different, but we can always adjust the other parameters accordingly. We set the promotion threshold to $h = 40$.

To show that introducing social recommendation via the Friends interface lets stories with smaller r get promoted to the front page, we first obtain an analytic solution for the maximum number of votes a story can receive on the upcoming stories queue without the social network effect. We set $v_f = v_s = v_m = 0$ and convert Equation 7 to a differential form by taking $\Delta t \to 0$:

$$\frac{dm}{dt} = rcc_u^{k_u t}N \tag{8}$$

This equation's solution is

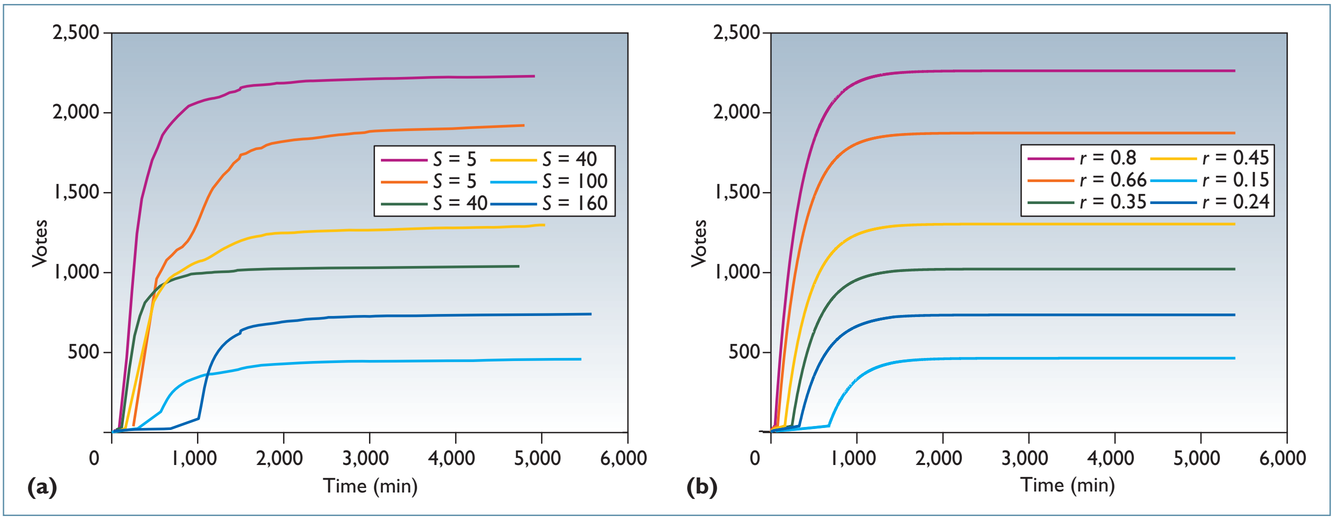


Figure 9. Comparison of data and analytic predictions. (a) Evolution of the number of votes received by six stories from the May data set; S denotes the submitter's reverse friends. (b) Predictions of the model for the same S values.

$$m(T) = rcN\left(c_u^{k_u T} - 1\right) / k \log c_u + 1.$$

Because $c_u < 1$, the exponential term will vanish for large times and leave us with $m(T \to \infty) = -rcN/(k_u \log c_u) + 1 \approx 42r + 1$. Hence, the maximum rating a story can receive on only the upcoming pages is 43. Because the threshold on Digg appears to be set around this value, no story can get promoted to the front page without other effects, such as users reading stories through the Friends interface.

Suppose the Friends interface lets users read only the stories their friends submit. Figure 8 shows how the ratings of three stories with $r = 0.1$, $r = 0.5$, and $r = 0.9$ change over time. For the chosen parameter values, a story posted by an unknown user ($S = 0$) never gathers enough votes to exceed the promotion threshold h. Even a highly interesting story with $r = 0.9$ languishes in the upcoming queue until it eventually disappears. A story posted by a user with $S = 80$ will be promoted to the front page if it's interesting enough — for example with $r \geq 0.5$ (Figure 8b). The more interesting story is promoted faster than a less interesting story, a general feature of collective voting. Stories posted by better-connected users follow the same pattern, although the interestingness value a story needs to get promoted is smaller — that is, a story with $r = 0.1$ posted by a user with $S = 400$ will be promoted.

Figure 9a shows the evolution of the number of votes received by six real stories from the Digg data set; S denotes the number of reverse friends the story's submitter had at submission time. Figure 9b shows solutions from Equation 7 for the same S values and different r values, chosen to produce best agreement with the data. Overall, qualitative agreement exists between the data and the model, indicating that the Digg user interface's basic features are enough to explain collaborative rating patterns. The only significant difference between the data and the model is visible in the lower two lines. In the data, a story posted by the user with $S = 100$ is promoted before the story posted by the user with $S = 160$, but saturates at a smaller value of diggs than the latter story. In the model, the story with bigger r is promoted first and gets more diggs. This disagreement isn't too surprising, given the number of approximations I made while constructing the model (I'll discuss modeling's limitations in a later section). For example, I assumed that the combined social network of voters grows at the same rate for all stories, which can't be true. If the combined social network grew at a slower rate for the story posted by the user with $S = 160$, it would explain the delay in promotion to the front page. Another effect I didn't consider is that a story could have a different r for users within the submitter's social network than for the general Digg audience. We can, however, extend the model to include inhomogeneous r.

Modeling as a Design Tool

Designing a complex system such as Digg, which exploits the emergent behavior of many independent evaluators, is exceedingly difficult. The choices made in the user interface — for example, whether to let users see the stories their friends

Previous Research in Social Filtering

Online networks' proliferation[1] has generated interesting data sets about the behavior of large groups "in the wild." Early studies focused on collecting social network information from citation,[2] coauthorship,[3] and email data.[4] Social media's rise has introduced yet another interesting domain in which to study the collective behavior of large numbers of connected individuals. Researchers are investigating various topics, from detecting[5] and influencing[6,7] trends in public opinion to tagging systems' evolution.[8–10]

Many Web sites that provide information (or sell products or services) use collaborative filtering (CF) technology to suggest relevant documents (or products and services) to their users. Amazon and Netflix, for example, use collaborative filtering to recommend new books or movies, respectively. CF-based recommendation systems[11] try to find users with similar interests by asking them to rate products and then comparing those ratings. Researchers in the past have recognized that social networks present in the recommender system's user base can be induced from the explicit and implicit declarations of user interest, and that the system can in turn use these networks to make new recommendations.[12,13] Social media sites such as Digg[14] and Flickr[15] are, to the best of my knowledge, the first systems to let users explicitly construct social networks and use them to get personalized recommendations.

Social navigation, a concept closely linked to CF, helps users evaluate information quality or guides them to new information sources by exposing information about the choices other users have made. Social navigation works "through information traces left by previous users for current users,"[16] much like footprints in the snow help guide pedestrians through a featureless snowy terrain. Exposing this information — also called social influence — affects collective decision making[17] and leads to a large variance in popularity for items of similar quality. Unlike the main article, these research projects took into account global information about others' preferences (similar to best-seller lists and top-ten albums), not choices others made within a user's own community. We believe that exposing local information about others' choices can lead to more effective collective decision making.

Fang Wu and Bernardo Huberman[18] recently studied the dynamics of collective attention on Digg. They proposed a simple stochastic model, parametrized by a single quantity that characterizes the rate at which interest in a news article decays. They collected data about how the number of votes that a front-page story received evolved over one month and showed that vote distribution can be described mathematically. They found that interest in a story peaks when the story first hits the front page and then decays with time, with a half-life of roughly one day, which corresponds to the average length of time a story spends on page one of the front page. The problem Wu and Huberman studied is complementary to the one described in the main text — they studied stories' dynamics after they hit the front page, but didn't identify a mechanism for the spread of interest in stories. On the other hand, I propose social networks as a mechanism for spreading a story's visibility and model vote evolution both before and after the story hits the front page. The novelty parameter in their model seems related to a combination of visibility and interestingness parameters in my model, and readers should view their model as an alternative.

The main text borrows techniques from mathematical analysis of multiagent systems' collective behavior. My team's earlier work proposed a formal framework for creating mathematical models of collective behavior in groups of multiagent systems.[19] We also successfully applied this framework to study collective behavior in robot groups.[20–22]

continued on p. 27

liked or the most popular stories within the last week or month — can dramatically affect system behavior as a whole. The designer must also consider the trade-offs between story timeliness and interestingness, how often stories are promoted, and the promotion algorithm itself. As I described earlier, Digg's old promotion algorithm alienated many users by making them feel that top users controlled the front page. Changes to the algorithm appeared to alleviate some of these concerns (while perhaps creating new ones). Unfortunately, few tools exist, short of running the system, that let developers explore different system designs.

Mathematical modeling and analysis can help us explore the design space of collaborative rating algorithms, despite possible limitations. I demonstrated previously that a story with low r posted by a well-connected user will be promoted to the front page. If the developer wants to prevent uninteresting stories from getting to the front page, he or she could change the promotion algorithm to make it difficult for people with bigger social networks to get their stories promoted. For instance, the promotion threshold could be set to be a function of S.

Modeling Limitations

I made several assumptions and abstractions while constructing the mathematical model and choosing its parameters. Some of the assumptions affected the model's structure — for example, the only terms that contribute to a story's visibility came from users viewing the front page or upcoming stories queue, or seeing the stories their friends recently submitted or dugg; I didn't include other browsing modalities in the model. In the Technol-

Previous Research in Social Filtering (cont.)

Although human behavior is, generally, far more complex, within the context of a collaborative rating system, Digg users show simple behaviors that can be analyzed mathematically. By comparing the analysis results with real-world data extracted from Digg, I show that mathematical modeling is a feasible approach for studying online users' collective behavior.

References

1. L. Garton, C. Haythornthwaite, and B. Wellman, "Studying Online Social Networks," *J. Computer Mediated Comm.*, vol. 3, no. 1, 1997; http://jcmc.indiana.edu/vol3/issue1/garton.html.
2. E. Otte and R. Rousseau, "Social Network Analysis: A Powerful Strategy, also for the Information Sciences," *J. Information Science*, vol. 28, no. 6, 2002, pp. 441–453.
3. M.E.J. Newman, "The Structure of Scientific Collaboration Networks," *Proc. Nat'l Academy of Sciences*, vol. 98, Nat'l Academy of Sciences, 2001, pp. 404–409.
4. H. Ebel, L.I. Mielsch, and S. Bornholdt, "Scale-Free Topology of Email Networks," *Physical Rev.*, vol. E 66, 2002, p. 035103R.
5. E. Adar et al., "Implicit Structure and the Dynamics of Blogspace," *Proc. Workshop on Weblogging Ecosystem, 13th Int'l World Wide Web Conf.*, ACM Press, 2004; www.hpl.hp.com/research/idl/papers/blogs/blogspace-draft.pdf.
6. P. Domingos and M. Richardson, "Mining the Network Value of Customers," *Proc. Knowledge Discovery in Databases* (KDD 01), ACM Press, 2001, pp. 57–66.
7. D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the Spread of Influence through a Social Network," *Proc. 9th ACM Int'l Conf. Knowledge Discovery and Data Mining* (SIGKDD 03), ACM Press, 2003, pp. 137–146.
8. S.A. Golder and B.A. Huberman, *The Structure of Collaborative Tagging Systems*, tech. report, HP Labs, 2005; www.hpl.hp.com/research/idl/papers/tags/.
9. P. Mika, "Ontologies Are Us: A Unified Model of Social Networks and Semantics," *Int'l Semantic Web Conf.* (ISWC 05), IEEE CS Press, 2005; www.cs.vu.nl/~pmika/research/papers/ISWC-folksonomy.pdf.
10. C. Marlow et al., "Ht06, Tagging Paper, Taxonomy, Flickr, Academic Article, to Read," *Proc. 17th Conf. Hypertext and Hypermedia* (Hypertext 06), ACM Press, 2006, pp. 31–40.
11. J.A. Konstan et al., "GroupLens: Applying Collaborative Filtering to Usenet News," *Comm.ACM*, vol. 40, no. 3, 1997, pp. 77–87.
12. H. Kautz, B. Selman, and M. Shah, "Referralweb: Combining Social Networks and Collaborative Filtering," *Comm. ACM*, vol. 4, no. 3, 1997, pp. 63–65.
13. S. Perugini, M. André Gonçalves, and E.A. Fox, "Recommender Systems Research: A Connection-Centric Survey," *J. Intelligent Information Systems*, vol. 23, no. 2, 2004, pp. 107–143.
14. K. Lerman, "Social Networks and Social Information Filtering on Digg," *Proc. Int'l Conf. Weblogs and Social Media* (ICWSM 07), ICWSM, 2007.
15. K. Lerman and L. Jones, "Social Browsing on Flickr," *Proc. Int'l Conf. Weblogs and Social Media* (ICWSM 07), ICWSM, 2007.
16. A. Dieberger et al., "Social Navigation: Techniques for Building More Usable Systems," *Interactions*, vol. 7, no. 6, 2000, pp. 36–45.
17. M.J. Salganik, P.S. Dodds, and D.J. Watts, "Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market," *Science*, vol. 311, no. 5762, 2006, p. 854.
18. F. Wu and B.A. Huberman, *Novelty and Collective Attention*, tech. report, Information Dynamics Laboratory, Hewlett-Packard Labs, 2007.
19. K. Lerman, A. Martinoli, and A. Galstyan, "A Review of Probabilistic Macroscopic Models for Swarm Robotic Systems," E. Sahin and W. Spears, eds., *Swarm Robotics Workshop: State-of-the-Art Survey*, LNCS 3342, Springer-Verlag, pp. 143–152.
20. K. Lerman and A. Galstyan, "Mathematical Model of Foraging in a Group of Robots: Effect of Interference," *Autonomous Robots*, vol. 13, no. 2, 2002, pp.127–141.
21. A. Martinoli, K. Easton, and W. Agassounon, "Modeling of Swarm Robotic Systems: A Case Study in Collaborative Distributed Manipulation," *Int'l J. Robotics Research*, vol. 23, no. 4, 2004, pp. 415–436.
22. K. Lerman et al., "Analysis of Dynamic Task Allocation in Multi-Robot Systems," *Int'l J. Robotics Research*, vol. 25, no. 3, 2006, pp. 225–242.

ogy section, for example, a user can choose to see only the stories that received the most votes during the preceding 24 hours (Top 24 Hours) or in the past 7, 30, or 365 days. In the model, I considered only the default Newly Popular browsing option, which shows the stories in the order they were promoted to the front page. I assume that most users choose this option. If data shows that other browsing options are popular, these terms can be included in the model to explain the observed behavior. Likewise, if users also choose to see the stories their friends have commented on, this option can be included from the Friends interface as well.

In addition to the model structure, I made several assumptions about the form of the terms and parameters. Although a large variance must exist in Digg user behavior, I chose to represent these behaviors by single valued parameters, rather than distributions. Thus, I assume a constant rate N at which users visit Digg. I also assume that a story's interestingness is the same for all users. In the future, I plan to explore how using parameter value distributions to describe the variance of user behavior affects the dynamics of collaborative rating.

The assumptions I made helped keep the model tractable, although a question remains whether I abstracted away any important factors, thus invalidating the model's results. I think the simple model presented here includes the salient features of Digg users' behavior. The model qualitatively explains observed collective voting patterns. If I need to quantitatively reproduce experimental data, or if I see a significant disagreement between the data and the model's predictions, I will need to include all browsing modalities and variance in user behavior.

The new social media sites offer a glimpse into the Web's future, where, rather than passively consuming information, users will actively participate in creating, evaluating, and disseminating it. Social media sites such as Digg show that it's possible to exploit others' activities to solve hard information-processing problems. We expect progress in this field to continue to bring novel solutions to problems in information processing, personalization, search, and discovery. ◨

Acknowledgments

This research is based on work supported in part by the US National Science Foundation under award numbers IIS-0535182, IIS-0413321, and BCS-0527725. I thank Fetch Technologies for providing wrapper-building and execution tools and Dipsy Kapoor for assistance with data processing.

References

1. J.A. Konstan et al., "GroupLens: Applying Collaborative Filtering to Usenet News," *Comm. ACM*, vol. 40, no. 3, 1997, pp. 77–87.
2. P. Domingos and M. Richardson, "Mining the Network Value of Customers," *Proc. Knowledge Discovery in Databases* (KDD 01), ACM Press, 2001, pp. 57–66.
3. J. Leskovec, L.A. Adamic, and B.A. Huberman, "The Dynamics of Viral Marketing," *Proc. 7th ACM Conf. Electronic Commerce* (EC 06), ACM Press, 2006, pp. 228–237.
4. D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the Spread of Influence through a Social Network," *Proc. 9th ACM Int'l Conf. Knowledge Discovery and Data Mining* (SIGKDD 03), ACM Press, 2003, pp. 137–146.
5. K. Lerman, "Social Networks and Social Information Filtering on Digg," *Proc. Int'l Conf. Weblogs and Social Media* (ICWSM 07), ICWSM, 2007.
6. K. Lerman and L. Jones, "Social Browsing on Flickr," *Proc. Int'l Conf. Weblogs and Social Media* (ICWSM 07), ICWSM, 2007.
7. L. Page et al., *The Pagerank Citation Ranking: Bringing Order to the Web*, tech. report, Stanford Digital Library Technologies Project, 1998.
8. J. Warren and J. Jurgensen, "The Wizards of Buzz," *Wall Street J.* online, Feb. 2007; http://online.wsj.com/public/article/SB117106531769704150-zpK10wf4CJOB4IKoJS5anuNoi6Y_20080209.html.
9. F. Wu and B.A. Huberman, *Novelty and Collective Attention*, tech. report, Information Dynamics Laboratory, Hewlett-Packard Labs, 2007.
10. A. Papoulis, *Probability and Statistics*, Prentice Hall, 1990.
11. K. Maney, "Wisdom of Crowds," *USA Today*, 12 Sept. 2006; www.usatoday.com/money/industries/technology/maney/2006-09-12-wisdom-of-crowds_x.htm.
12. K. Lerman, A. Plangrasopchok, and C. Wong, "Personalizing Results of Image Search on Flickr," *Proc. AAAI Workshop Intelligent Techniques for Web Personalization*, AAAI Press, 2007, pp. 65–75.

Kristina Lerman is a project leader at the University of Southern California Information Sciences Institute and a research assistant professor in computer science at USC. Her research interests include mathematical modeling and analysis of multiagent systems — including robots and social networks — semantic modeling of information sources for automatic information integration, and integrating these disparate research threads in the social Web. Lerman has a PhD in physics from the University of California, Santa Barbara. Her research is inspired by heavy use of Flickr and Digg, among other social media sites. Contact her at lerman@isi.edu.