Social Information Processing in Social News Aggregation

Kristina Lerman
University of Southern California
Information Sciences Institute
4676 Admiralty Way
Marina del Rey, California 90292
lerman@isi.edu

June 28, 2007

Abstract

The rise of social media sites — blogs, wikis, and Digg — underscores the transformation of the Web to a participatory medium in which users are collaboratively creating, evaluating and distributing information. The innovations introduced by social media have lead to a new paradigm for interacting with information: *social information processing*. We study how the social news aggregator Digg exploits social information processing to solve the problems of document recommendation and rating. First, we show that social networks play an important role in document recommendation. The second contribution of this paper consists of a mathematical model that describes how collaborative evaluation of documents emerges from the independent decisions made by many users. The model takes into account users behavior: e.g., whether they are reading stories on the front page or through a Friends interface. Solutions of the model reproduce the observed ratings received by actual stories on Digg.

1 Introduction

The label *social media* has been attached to a quickly growing number of Web sites whose content is primarily user driven. Examples of such sites include the following: blogs (personal online journals that allow users to share their thoughts and receive feedback on them), Wikipedia (a collectively written and edited online encyclopedia), and Flickr, Del.icio.us, and Digg (Web sites that allow users to share, discuss, and rank photos, Web pages, and news stories respectively). Other sites (e.g., Amazon's Mechanical Turk) allow users to collaboratively find innovative solutions to hard problems. The rise of social media underscores a transformation of the Web as fundamental as its birth. Rather than simply searching for, and passively consuming, information, users are collaboratively creating, evaluating, and distributing information. In the near future, new information-processing applications enabled by social media will include tools for personalized information discovery, applications that exploit the "wisdom of crowds" (e.g., emergent semantics and collaborative information evaluation), deeper analysis of community structure to identify trends and experts, and many other still difficult to imagine.

Social media sites share four characteristics: (1) Users create or contribute content in a variety of media types; (2) Users annotate content with tags; (3) Users evaluate content, actively by voting or passively by using it; and (4) Users create social networks by designating other users with similar interests as contacts or friends. We believe that social media facilitate new ways of interacting with information and enhance collaborative problem solving through what we call *social information processing*.

In this paper, we study how the social news aggregator Digg[1] uses social information processing to solve two

[1] http://digg.com

long standing problems of document recommendation and rating. The functionality of Digg is very simple: users submit stories they find online, and other users rate these stories by voting. Digg also allows users to create social networks by adding other users as friends and provides an interface to easily track their activities: e.g., what stories users within their social network read and liked. Each day, Digg promotes a handful of stories to its front pages based on the stories' voting patterns. Therefore, the promotion mechanism does not depend on the decisions of a few editors, but emerges from the activities of many users. This type of collective decision making can be extremely effective in breaking news, often outperforming special-purpose algorithms. For example, the news of Rumsfeld's resignation in the wake of the 2006 U.S. Congressional elections broke Digg's front page within 3 minutes of submission and 20 minutes before it was related by Google News [27]. In addition to promoting news stories, Digg ranks users by how successful they are at getting their stories promoted to the front page.

The first contribution of this paper is an empirical study of how social networks are used to discover new interesting content. This type of *social filtering* or *social recommendation* is an effective alternative to collaborative filtering (CF), a popular recommendation technology used by commercial giants like Amazon and Netflix. CF-based recommender system asks users to express their opinions by rating items, and then suggests new items that were liked by other users with similar opinions. One noted problem with CF is that users are generally resistant to rating [10]. In contrast, on social media sites users express their tastes and preferences by creating personal social networks of tens to hundreds (even thousands) of friends. Social recommendation has been studied in the context of the spread of innovation and viral marketing [3, 17]. These studies allow advertisers to target their messages [9] in order to get the most out of the "word of mouth" effect. We are interested in the inverse problem: how social networks can be used to effectively filter the vast streams of information [11, 14].

Another outstanding problem in information processing is how to evaluate the quality of documents. This problem crops up daily in information retrieval and Web search, where the goal is to find, among the terabytes of data accessible online, the information that is most relevant to a user's query. The standard practice of search engines is to identify all documents using the terms that appear in a user's query, and rank the results according to their quality or importance. Google revolutionized Web search by exploiting the link structure of the Web, created through independent activities of many Web page authors, in order to evaluate the contents of information on Web pages [24]. Similarly, social news aggregators Digg and Reddit[2] rely on the opinions of their users to evaluate the quality of news stories.

The second contribution of this paper is a mathematical model that describes the dynamics of collaborative rating of the quality of news stories. The model takes into account social influence exerted by users through their social networks. We show that the model correctly predicts the observed behavior of ratings received by actual stories on Digg.

The paper is organized as follows. In Section 2, we describe Digg's functionality and features in greater detail. In Section 3, we study the dynamics of collaborative rating of news stories on Digg. We show in Section 3.1 that social networks have a strong impact on the number of votes received by a story through the mechanism of social filtering. In Section 4, we develop a mathematical model of collaborative rating and discuss its behavior. Although we validate our model on Digg, we argue that the results are general and that mathematical analysis can be used to guide the design of collaborative rating systems. Finally, in Section 4.5, we discuss limitations of mathematical modeling, and identify new directions for future research.

2 Anatomy of Digg

Digg is a social news aggregator that relies on users to submit stories and moderate them. A typical Digg page is shown in Figure 1. When a story is submitted, it goes to the upcoming stories queue. There are 1-2 new submissions every minute and they are displayed in reverse chronological order of being submitted, 15 stories to a page, with the most recent story at the top. The story's title is a link to the source, while clicking

[2] http://reddit.com

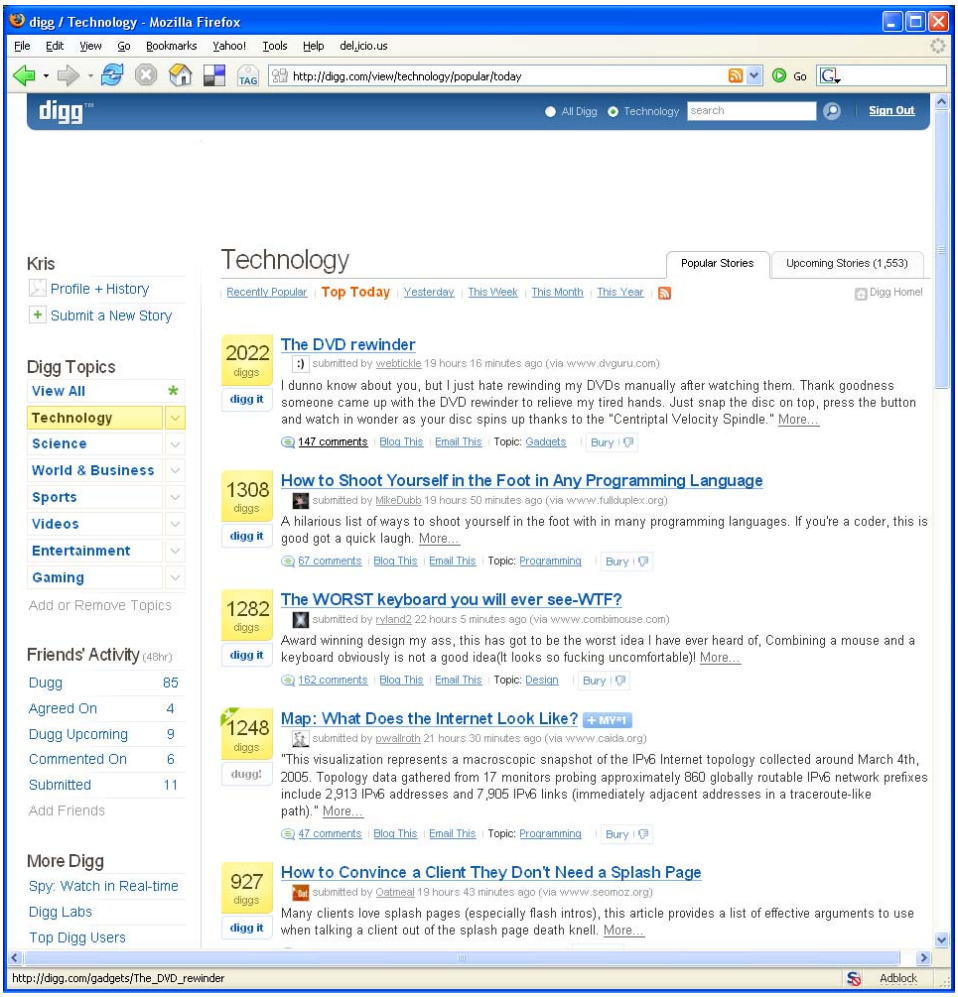


Figure 1: Digg front page showing the technology section

on the number of diggs (votes) the story received takes one to the page describing the story's activity on Digg: the discussion around it, the list of people who voted on it, etc.

A user votes on a story by "digging" it. Digging a story also saves it to user's history. Digg also allows users to "bury" stories that are determined to be spam, duplicates or contain inappropriate materials. "Burying" a story does not reduce its rating, as voting a story down on the social news aggregator Reddit does. Rather, if enough people have "buried" a story, it is permanently removed from Digg.

Emergent front page When a story gets enough votes, it is promoted to the front page. The vast majority of people who visit Digg daily, or subscribe to its RSS feeds, read only the front page stories; hence, getting to the front page greatly increases the story's visibility. Although the exact promotion mechanism is kept secret and changes periodically, it appears to take into account the number of votes the story receives. Digg's popularity is fueled in large part by the phenomenon of the emergent front page which is formed by consensus between many independent users.

Other social media sites rely on similar mechanisms to showcase select content. Every day the photo sharing site Flickr[3] chooses 500 most "interesting" of the newly uploaded images to feature on its Explore page. The

[3]http://flickr.com

selection algorithm takes into account how many times the image was viewed, commented on it or marked as a favorite.[4] Therefore, Flickr's Explore page also arises from decisions made by many users. Similarly, the social bookmarking site Delicious[5] showcases the most popular of the recently tagged Web pages.

Social networks Digg allows users to designate others as friends and makes it easy to track their activities. The Friends interface in the left column of the page in Figure 1 summarizes the number of stories friends have submitted, commented on or dugg recently. Tracking activities of friends is a common feature of many social media sites and is one of their major draws. It 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 put others to the task of finding and filtering information for them — what we call *social filtering*.

Top users Until February 2007 Digg ranked users according to how many of the user's stories were promoted to the front page. User ranked number one had submitted the most front page stories; user ranked number two had fewer stories promoted to the front page, and so on. Clicking on the Top Users link allowed one to browse through the ranked list of users. There is speculation that ranking users increased competition, leading some users to be more active in order to improve their ranking. Digg discontinued making the list of top users publicly available, citing concerns that marketers were paying top users to promote their products and services [30].

3 Social filtering

We tracked both upcoming and front page stories in Digg's technology section by scraping Digg site with the help of Web wrappers, created using tools provided by Fetch Technologies[6]:

digg-frontpage wrapper extracts a list of stories from the first 14 front pages. For each story, it extracts submitter's name, story title, time submitted, number of votes and comments the story received, along with the names of the first 216 users who voted on the story.

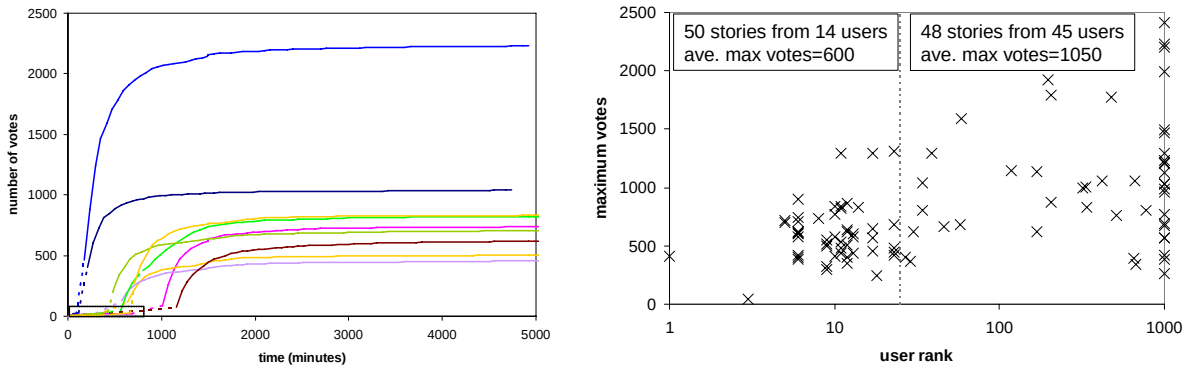
digg-all wrapper extracts a list of stories from the first 20 pages in the upcoming stories queue. For each story, it extracts the submitter's name, story title, time submitted, number of votes and comments the story received.

top-users wrapper extracts information about the top 1020 of the recently active users. For each user, it extracts the number of stories that user has submitted, commented and voted on; number of stories that have been promoted to the front page; user's rank; the list of friends, as well as reverse friends or "people who have befriended this user."

Digg-frontpage and *digg-all* wrappers were executed hourly over a period of a week on several occasions. *Top-users* wrapper was executed weekly starting in July 2006 to gather snapshots of the social network of the top Digg users.

We identified stories that were submitted to Digg over the course of approximately one day and followed them over several days. Of the 2858 stories submitted by 1570 distinct users over this time period, only 98 stories by 60 users made it to the front page. Figure 2(a) shows evolution of the ratings (number of votes) of select stories. The basic dynamics of all stories appears the same. While in the upcoming queue, a story accrues votes at some slow rate. Once it is promoted to the front page, it accumulates votes at a much faster rate. As the story ages, accumulation of new votes slows down [31], and the story's rating saturates

[4] http://flickr.com/explore/interesting/
[5] http://del.icio.us
[6] http://fetch.com/



(a) (b)

Figure 2: (a) Dynamics of votes on select stories over a period of four days. The small rectangle indicates the time the stories where in the upcoming stories queue, while dashes indicate transition to the front page. (b) Maximum votes received by stories during the period of observation vs submitter's rank. Symbols on the right axis correspond to low-rated users with rank> 1020.

at some value, which we call "interestingness", which indicates how interesting the story is to the general Digg community.

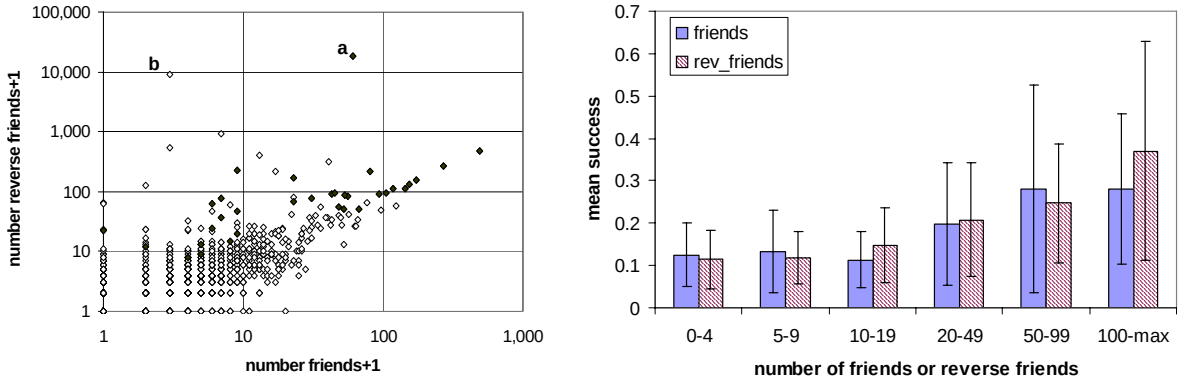
It is worth noting that the top-ranked users are not submitting stories that get the most votes. This is shown graphically in Figure 2(b), which displays the maximum number of votes received by stories vs submitter's rank. Slightly more than half of the stories came from 14 top-ranked users (rank< 25), and 48 from 45 low-ranked users. The average "interestingness" of the stories submitted by the top-ranked users is 600, almost half the average "interestingness" of the stories submitted by low-ranked users. Top-ranked users are also 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 1020 users as of May 2006, the top 3% of the users were responsible for 35% of the stories.

3.1 Social networks and recommendation

If top-ranked users are not submitting the most interesting stories, why are they so successful? We believe that social filtering plays a role in helping promote stories to the front page. As explained above, Digg allows users to track friends' activities: the stories they have submitted, commented and voted on. *We believe that users employ the Friends interface to filter the tremendous number of new submissions on Digg to find new interesting stories.*

Note that the friend relationship is asymmetric. When user A lists user B as a *friend*, A is able to watch the activities of B but not vice versa. We call A the *reverse friend* of B. Figure 3(a) shows the scatter plot of the number of friends vs reverse friends of the top 1020 Digg users as of May 2006. Black symbols correspond to the top 33 users. For the most part, users appear to take advantage of Digg's social networking feature, with the top users having bigger social networks. Two of the biggest celebrities (watched by most people) are users marked a and b on Figure 3(a), corresponding *kevinrose* and *diggnation*, respectively, one of the founders of Digg and a podcast of the popular Digg stories.

First, we present indirect evidence for social filtering on Digg by showing that user's success is correlated with his social network size. A user's success rate is defined as the fraction of the stories the user submitted that have been promoted to the front page. We use the statistics about the activities of the top 1020 users to show that users with bigger social networks are more successful at getting their stories promoted to the front page. We only include users who have submitted 50 or more stories (total of 514 users). The correlation between users's mean success rate and the size of their social network is shown in Figure 3(b). Data was



(a) (b)

Figure 3: (a) Scatter plot of the number of friends vs reverse friends for the top 1020 Digg users. (b) Strength of the linear correlation coefficient between user's success rate and the number of friends and reverse friends he has. story

binned to improve statistics. Despite large error bars, there is a significant correlation between users's success rate and the size of their social network, more importantly, the number of reverse friends they have.

In the sections below we present additional evidence that the Friends interface is used to find new interesting stories. We show this by analyzing two sub-claims: (a) *users digg stories their friends submit*, and (b) *users digg stories their friends digg*. By "digging" the story, we mean that users like the story and vote on it.

3.1.1 Users digg stories their friends submit

In order to show that users digg stories their friends submit, we used *digg-frontpage* wrapper to collect 195 front page stories, each with a list of the first 216 users to vote on the story (15,742 distinct users in total). The name of the submitter is first on the list.

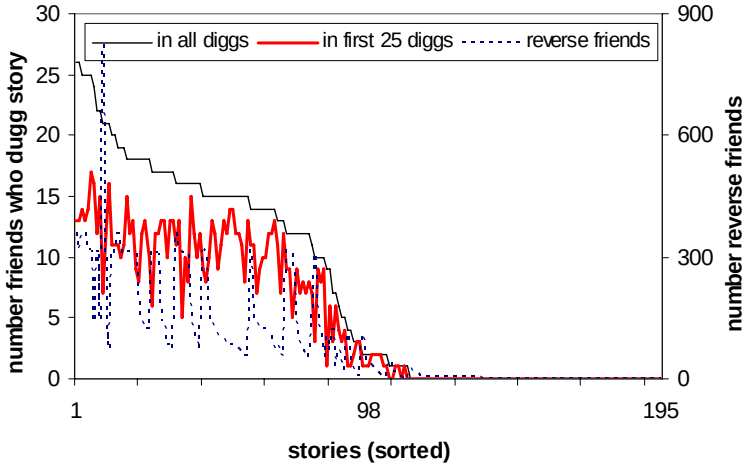
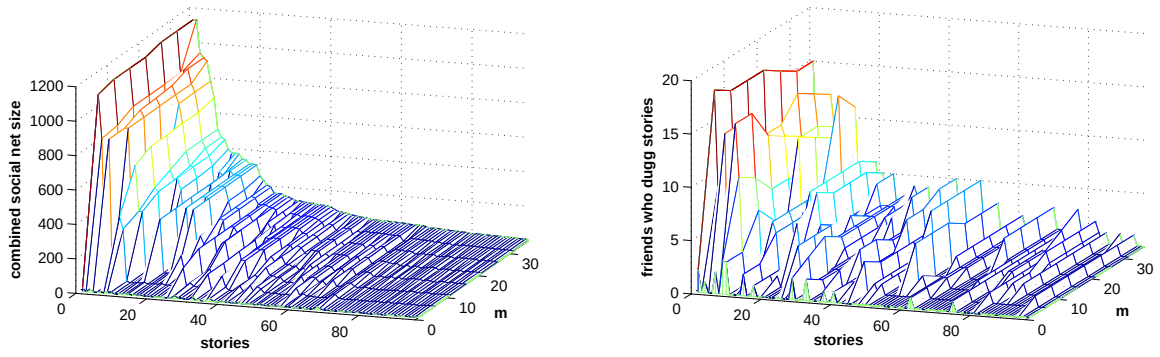


Figure 4: Number of voters who are also among the reverse friends of the user who submitted the story

We can compare this list, or any portion of it, with the list of the reverse friends of the submitter. The dashed line in Figure 4 shows the size of the social network (number of reverse friends) of the submitter. More than half of the stories (99) were submitted by users with more than 20 reverse friends, and the rest by



(a) (b)

Figure 5: (a) Number of reverse friends of the first m voters for the stories submitted by poorly-connected users. (b) Number of friends of the first m voters who also voted on the stories.

	diggers	m=1	m=6	m=16	m=26	m=36
(a)	visible to friends	34	75	94	96	96
(b)	dugg by friends	10	23	37	46	49
(c)	probability	0.005	0.028	0.060	0.077	0.090

Table 1: Number of stories posted by poorly-connected users that were (a) made visible to others by digging activities of well-connected users, (b) dugg by friends of the first m diggers within the next 25 diggs, and for the stories that were dugg by friends, (c) the average probability that the observed numbers of friends dugg the story by chance

poorly connected users.[7] The thin line shows the number of voters who are also among the reverse friends of the submitter. All but two of the stories (submitted by users with 47 and 28 reverse friends) were dugg by submitter's reverse friends.

We use simple combinatorics [25] to compute the probability that k of 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 that came 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 dugg by at least one friend) that the observed numbers of reverse friends voted on the story by chance is $P = 0.005$, making it highly unlikely.[8] Moreover, users digg stories submitted by their friends very quickly. The heavy red line in Figure 4 shows the number of reverse friends who were among the first 25 voters. The probability that these numbers could have been observed by chance is even less — $P = 0.003$. We conclude that users digg — or tend to like — the stories their friends submit. As a side effect, by enabling users to quickly digg stories submitted by friends, social networks play an important role in promoting stories to the front page.

3.1.2 Users digg stories their friends digg

Do social networks also help users discover interesting stories that were submitted by poorly-connected users? Digg's Friends interface allows users to see the stories their friends have liked (dugg). As well-connected users digg stories submitted by users who have few or no reverse friends, are others within his or her social network more likely to read them?

[7] These users have rank > 1020 and were not listed as friends of any of the 1020 users in our dataset. It is possible, though unlikely, that they have reverse friends.

[8] If we include in the average the two stories that were not dugg by any of the submitter's friends, we end up with a higher, but still significant P=0.023.

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

At the time of submission ($m = 1$), only 34 of the 96 stories were visible to others within the submitter's social network and ten of these were dugg by submitter's reverse friends within the first 25 votes. After fifteen more users have voted, almost all stories are now visible through the Friends interface. Table 1 summarizes the observations and presents 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 above the 0.05 significance level, possibly reflecting story's increased visibility on the front page. Although the effect is not quite as dramatic as one in the previous section, we 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.

3.2 Changing the promotion algorithm

Digg's goal is to feature only the most interesting of the submitted stories on its front page, and it employs aggregated opinion of thousands of users, rather than a few dedicated editors, to achieve this goal. We showed above that social networks play an important role in social filtering and recommendation. Since some users are more active than others, direct implementation of social filtering may lead to "tyranny of the minority," where a lion's share of front page stories come from users with the most active social networks.

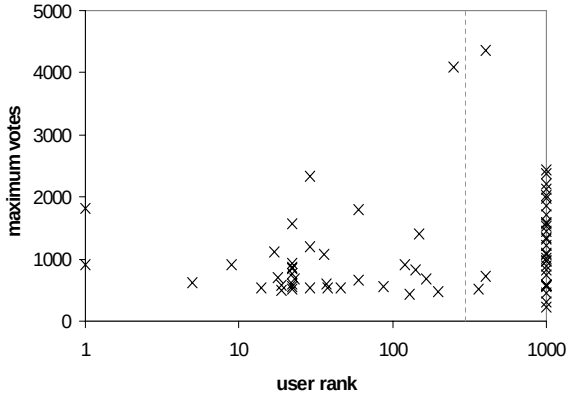


Figure 6: Maximum number of votes received by front page stories vs submitter's rank. Data was collected from stories submitted to Digg in early November 2006, after the change in the promotion algorithm. The vertical line divides the set in half. Symbols on the right hand axis correspond to low-rated users with rank$>$ 1020.

A similar finding [7] in September 2006 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 [18] prompted Digg to change the algorithm it uses to promote stories. In order to discourage what was seen as "bloc voting," the new algorithm looked "at the unique digging diversity of the individuals digging the story" [28]. Analysis of the votes received by stories submitted in early November 2006 indicates that the algorithm change did achieve the desired effect of reducing the top user dominance on the front page. Analysis shows that of the 3072 stories submitted by 1865 users over a period of about a day, 71 stories by 63 users were promoted to the front page. Figure 6 shows the maximum number of diggs received by these stories over a six day period vs submitter's rank. Compared to the May data (Figure 2(b)), the front page now has a greater diversity of users, with fewer users responsible for multiple front page stories (1.2 stories/submitter compared to 1.6 stories/submitter). Rank distribution is less skewed towards top-ranked users than before: half of the stories

came from users with rank< 300, rather than rank< 25 in the May dataset. There is also a smaller spread in the mean interestingness of stories submitted by top- and low-ranked users (960 vs 1270 in November *cf* 600 vs 1050 in May).[9]

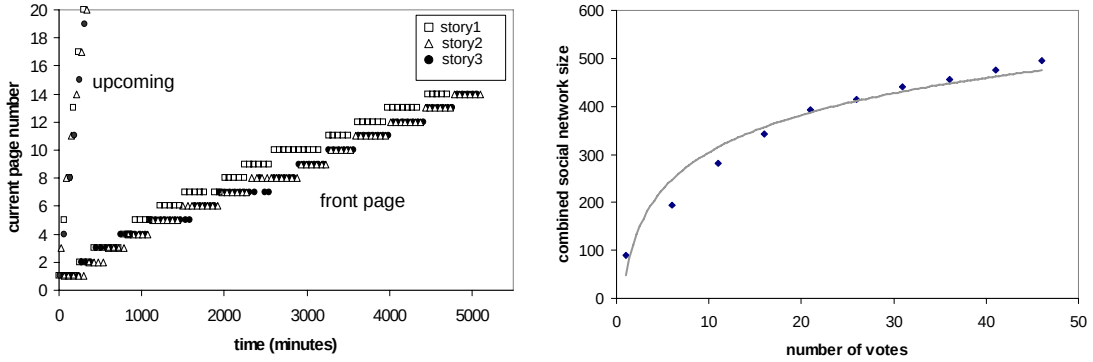
Although these changes may be seen as a positive development, the promotion algorithm changes may have had unintended consequences: e.g., discouraging users from joining social networks because their votes will be discounted. Mathematical analysis, described in the sections below, can be used as a tool to investigate the consequences of changes in the promotion algorithm. Rather than being a liability, however, social networks are a useful feature of social media sites, which can be used to to personalize and tailor information to individual users [16], and drive the development of new *social search algorithms*. Digg can offer personalized front pages to every user, selected from their friends' submission and voting history.

4 Mathematical model of collaborative rating

In this section we present a mathematical model that describes how the number of votes received by a story changes in time. Our goal is not only to produce a model that can explain — and predict — how the front page emerges on Digg, but can also be used as a tool to study the behavior of different collaborative rating algorithms.

We parameterize a story by its *interestingness* coefficient r, which gives the probability that a story will receive a positive vote once seen. The number of votes a story receives depends on its *visibility*, which simply means how many people can see and follow the link to the story. The following factors contribute to a story's visibility:

- visibility on the front page
- visibility in the upcoming stories queue
- visibility through the Friends interface



(a) (b)

Figure 7: (a) Current page number of a story on the upcoming stories queue and the front page vs time for three different stories. (b) Growth of the combined social network of the first 46 users to vote on a story

4.1 Visibility on Digg's pages

A story's visibility on the front page decreases as newly promoted stories push it farther down the list. While we do not have data about Digg visitors' behavior, specifically, how many proceed to page 2, 3 and so on,

[9]The overall increase in the maximum number of votes received by stories could reflect the growth of the Digg user base.

we propose to describe it by a simple model that holds that some fraction c_f of the 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 it is pushed down the list by the newer submissions. If a fraction c of Digg visitors proceed to the upcoming stories section, and of these, a fraction c_u proceed to the next upcoming page, then $cc_u N$ of Digg visitors see second page stories, and $cc_u^{q-1} N$ users see page q stories.

Figure 7(a) shows how the current page number, on the upcoming stories and the front page, changes in time for three randomly chosen stories from the May dataset. The change in a story's current page number can be fit by lines $q, p = k_{u,f} t$ with slopes $k_u = 0.060$ pages/m (3.60 pages/hr) on the upcoming stories and $k_f = 0.003$ pages/m (0.18 pages/hr) on the front page.

We use a simple threshold to model how a story is promoted 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 is greater than h, it is visible on the front pages. This seems to approximate Digg's promotion algorithm as of May 2006, since in our dataset we did not see any front page stories with fewer than 44 votes, nor did we see any upcoming stories with more than 42 votes.

4.2 Visibility through the Friends interface

The Friends interface offers the user ability to see the stories his friends have (i) submitted, (ii) liked (dugg), (iii) commented on during the preceding 48 hours or (iv) friends' stories that are still in the upcoming stories queue. Although it is likely that users are taking advantage of all four features, we will consider only the first two in the analysis. These closely approximate the functionality offered by other social media sites: e.g., Flickr allows users to see the latest images his friends uploaded, as well as the images a friend liked (marked as favorite). We believe that these features are more familiar to the user and used more frequently than the other features.

Friends of the submitter 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 these users visit Digg daily, and since they are 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 $v_s = a\Theta(S - at)\Theta(48 - t)$.[10] 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.

Friends of the voters As the story is dugg, it becomes visible to more users through the "see the stories my friends dugg" part of the Friends interface. Figure 7(b) 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, it's average value has consistent growth: $S_m = 112.0 * log(m) + 47.0$. Therefore, the story's visibility through the combined social network of the first m users who vote on it is $v_m = bS_m\Theta(h - m)\Theta(48hrs - t)$, where b is a scaling factor that depends on the length of the time interval: for hourly counts, it is $b = 1/24$.

[10] $\Theta(x)$ is a step function whose value is 1 when $x \geq 0$ and 0 when $x < 0$.

4.3 Dynamical model

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

$$v_f = c_f^{p(t)-1} N\Theta(m(t)-h) \quad (1)$$
$$v_u = cc_u^{q(t)-1} N\Theta(h-m(t))\Theta(24hrs-t) \quad (2)$$
$$v_s = a\Theta(S-at)\Theta(48hrs-t) \quad (3)$$
$$v_m = bS_m\Theta(h-m(t))\Theta(48hrs-t) \quad (4)$$

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 is visible in the upcoming stories pages; and 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, while step functions in v_s and v_m model the fact that it is 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:

$$p(t) = (k_f(t-T_h)+1)\Theta(T_h-t) \quad (5)$$
$$q(t) = k_u t + 1 \quad (6)$$

with $k_u = 0.060$ pages/min and $k_f = 0.003$ pages/min. T_h is the time the story is promoted to the front page.

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

$$\Delta m(t) = r(v_f + v_u + v_s + v_m)\Delta t \quad (7)$$

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

We show that introducing social recommendation via the Friends interface allows stories with smaller r to be promoted to the front page. First, we 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 $\Delta t \to 0$:

$$\frac{dm}{dt} = rcc_u^{k_u t} N \quad (8)$$

The solution of the above equation is $m(T) = rcN(c_u^{k_u T} - 1)/(k \log c_u) + 1$. Since $c_u < 1$, 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 the upcoming pages only is 43. Since the threshold on Digg appears to be set around this value, no story can be promoted to the front page without other effects, such as users reading stories through the Friends interface.

Suppose the Friends interface only allows users to read 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 in 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.

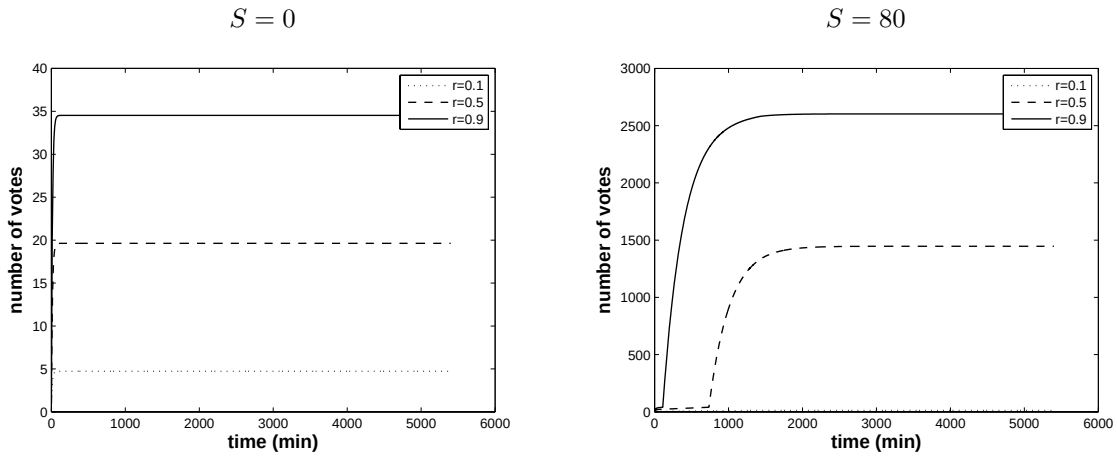


Figure 8: Effect of the submitter's social network on the evolution of a story's rating. Votes received by a story posted by (a) an unknown user with $S = 0$ and a (b) connected user with $S = 80$.

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 is interesting enough, e.g., with $r \geq 0.5$ (Figure 8(b)). 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 be promoted is smaller: e.g., a story with $r = 0.1$ posted by a user with $S = 400$ will be promoted.

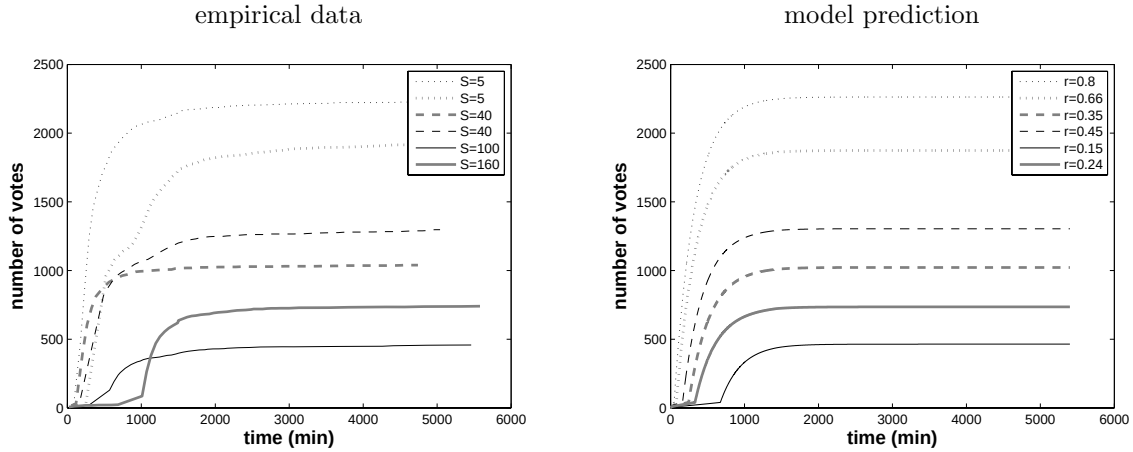


Figure 9: (a) Evolution of the number of votes received by six stories from the May dataset. S gives the number of submitter's reverse friends. (b) Predictions of the model for the same values of S.

Figure 9(a) shows the evolution of the number of votes received by six real stories from the Digg dataset. S denotes the number of reverse friends the story's submitter had at the time of submission. Figure 9(b) shows solutions of Equation 7 for the same values of S and different values of r. Overall there is qualitative agreement between the data and the model, indicating that the basic features of the Digg user interface we considered are enough to explain the patterns of collaborative rating. 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 smaller value of diggs than the latter story. In the model, the story with bigger r is promoted first and gets more

diggs. The disagreement is not too surprising, given the number of approximations made in the course of constructing the model (see Section 4.5 for discussion of modeling limitations). For example, we assumed that the combined social network of voters grows at the same rate for all stories, which cannot be true. If the combined social network grew at a slower rate for the story posted by user with $S = 160$, this would explain the delay in promotion to the front page. Another effect not currently taken into consideration is that a story could have a different r to users within the submitter's social network than to the general Digg audience. The model can be extended to include inhomogeneous r.

4.4 Modeling as a design tool

Designing a complex system like Digg, which exploits the emergent behavior of many independent evaluators, is exceedingly difficult. The choices made in the user interface, for example, whether to allow users to see the stories their friends liked or the most popular stories within the last week or month, can have a dramatic impact on the behavior of system as a whole. The designer has to also consider the tradeoffs between story timeliness and interestingness, how often stories are promoted, and the promotion algorithm itself. As described in Section 3.2, Digg's old promotion algorithm alienated many users by making them feel that top users controlled the front page. Changes to the promotion algorithm in November 2006 appeared to alleviate some of these concerns (while perhaps creating new ones). Unfortunately, there are few tools, short of running the system, that allow developers to explore different system designs.

We believe that mathematical modeling and analysis can be a valuable tool for exploring the design space of collaborative rating algorithms, despite the limitations described in Section 4.5. We saw above that a story with low r posted by a well connected user will be promoted to the front page. If it is desirable to prevent uninteresting stories from getting to the front page, the promotion algorithm could be changed to make it more difficult for people with bigger social networks to get their stories promoted. For example, the promotion threshold could be set to be a function of S.

4.5 Limitations of modeling

A number of assumptions and abstractions have been made in the course of constructing the mathematical model and choosing its parameters. Some of our assumptions affect the structure of the model: e.g., the only terms that contribute to the visibility of the story come from users viewing the front page, upcoming stories queue or seeing the stories one's friends have recently submitted or dugg, while other browsing modalities were not included in the model. In the Technology 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, we only considered the default "Newly popular" browsing option, which shows the stories in the order they have been promoted to the front page. We 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 of the Friends interface can also be included in our model.

In addition to the model structure, we made a number of assumptions about the form of the terms and the parameters. Although there must exist a large variance in Digg user behavior, we chose to represent these behaviors by single valued parameters, not distributions. Thus, we assume a constant rate users visit Digg, characterized by N in the model. We also assume that a story's interestingness is the same for all users. In future we plan to explore how using distributions of parameter values to describe the variance of user behavior affects the dynamics of collaborative rating.

The assumptions we make help keep the model tractable, although a question remains whether any important factors have been abstracted away so as to invalidate the results of the model. We claim that the simple model we present in the paper do include the salient features of the Digg users' behavior. We showed that the model qualitatively explain the observed collective voting patterns. If we need to quantitatively reproduce

experimental data, or see a significant disagreement between the data and predictions of the model, we will need to include all browsing modalities and variance in user behavior.

5 Previous research

The proliferation of online networks [5] has provided interesting datasets about the behavior of large groups in the wild. The early studies focused on collecting social network information from citation [23], co-authorship [22] and email [4] data. The rise of social media has introduced yet another interesting domain for the study of collective behavior of large numbers of connected individuals. Researchers are investigating a variety of topics, from detecting [1] and influencing [3, 9] trends in public opinion, to the evolution of tagging systems [6, 21, 19]. The focus of our research, on the other hand, is the role of social networks in information processing [11, 14].

Many Web sites that provide information (or sell products or services) use collaborative filtering technology to suggest relevant documents (or products and services) to its users. Amazon and Netflix, for example, use collaborative filtering to recommend new books or movies to its users. Collaborative filtering-based recommendation systems [10] try to find users with similar interests by asking them to rate products and then compare ratings to find users with similar opinions. Researchers in the past have recognized that social networks present in the user base of the recommender system can be induced from the explicit and implicit declarations of user interest, and that these social networks can in turn be used to make new recommendations [8, 26]. Social media sites, such as Digg, are to the best of our knowledge the first systems to allow users to explicitly construct social networks and use them for getting personalized recommendations. Unlike collaborative filtering research, the topic of this paper was not recommendation per se, but how social network-based recommendation affects the global rating of quality of information.

Social navigation, a concept closely linked to collaborative filtering, helps users evaluate the quality of information, or guide them to new information sources, by exposing information about the choices made by other users. Social navigation works "through information traces left by previous users for current users" [2] much like footprints in the snow help guide pedestrians through a featureless snowy terrain and pheromone trails left by ants help guide them to food sources. Exposing information about the choices made by others — social influence — has been has been shown [29] to affect collective decision making and lead to a large variance in popularity of similar quality items. Unlike the present work, these research projects took into account the global information about the preferences of others (similarly to the best seller lists and Top Ten albums), not choices made by others within a user's own community. We believe that exposing local information about the choices of others similar to you can lead to more effective collective decision making.

Wu and Huberman [31] have recently studied the dynamics of collective attention on Digg. They proposed a simple stochastic model, parametrized by a single quantity that characterizes the rate of decay of interest in a news article. They collected data about the evolution of the number of votes received by front page stories over a period of one month, and showed that the distribution of votes can be described by the model. 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 about a day, corresponding to the average length of time a story spends on page one of the front page. The problem studied by Wu and Huberman is complementary to ours. They studied dynamics of stories *after* they hit the front page. The authors did not identify a mechanism for the spread of interest in stories. On the other hand, we propose social networks as a mechanism for spreading a story's visibility and model evolution of votes both before and after the story hits the front page. The novelty parameter in their model seems to be related to a combination of visibility and interestingness parameters in our model, and their model should be viewed as an alternative.

This paper borrows techniques from mathematical analysis of collective behavior of multi-agent systems. Our earlier work proposed a formal framework for creating mathematical models of collective behavior in groups of multi-agent systems [15]. This framework was successfully applied to study collective behavior in groups of robots [12, 20, 13]. Although the behavior of humans is, in general, far more complex than the

behavior of robots, within the context of a collaborative rating system, Digg users show simple behaviors that can be analyzed mathematically. By comparing results of analysis with real world data extracted from Digg, we showed that mathematical modeling is a feasible approach to study collective behavior of online users.

6 Conclusion

The new social media sites offer a glimpse into the future of the Web, where, rather than passively consuming information, users will actively participate in creating, evaluating, and disseminating information. One novel feature of these sites is that they allow users to create personal social networks so as to easily keep track of friends' activities. Another novel feature is the collaborative evaluation of content, either explicitly through voting or implicitly through using content. Together, these innovations lead to a new paradigm for interacting with information, what we call *social information processing*. Social information processing enables users to collectively solve such hard information processing problems, such as document recommendation and filtering, and evaluating the quality of documents.

We studied social information processing on the social news aggregator Digg. We showed that social networks form a basis for an effective *social recommendation* system, suggesting to users the stories his friends have found interesting. Next, we studied collaborative rating, through voting, of news stories on Digg, focusing on the process by which consensus emerges from the distributed opinions of many voters. We created a mathematical model of the dynamics of collective voting and found that solutions of the model qualitatively agreed with the evolution of votes received by actual stories on Digg.

Besides offering a qualitative explanation of user behavior, mathematical modeling can be used as a tool to explore the design space of user interfaces. The design of complex systems such as Digg that exploit emergent behavior of large numbers of users is notoriously difficult, and mathematical modeling can help to explore the design space before a particular collective voting algorithm is actually implemented.

Social media sites, such as Digg, show that it is possible to exploit the activities of others 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.

Acknowledgements

This research is based on work supported in part by the National Science Foundation under Award Nos. IIS-0535182, IIS-0413321 and BCS-0527725. We are grateful to Fetch Technologies for providing wrapper building and execution tools and to Dipsy Kapoor for assistance with data processing.

References

[1] E. Adar, L. Zhang, L. A. Adamic, and R. M. Lukose. Implicit structure and the dynamics of blogspace. In *Workshop on the Weblogging Ecosystem, 13th International World Wide Web Conference*, 2004.

[2] A. Dieberger, P. Dourish, K. Hk, P. Resnick, and A. Wexelblat. Social navigation: techniques for building more usable systems. *interactions*, 7(6):36 –45, Nov/Dec 2000.

[3] P. Domingos and M. Richardson. Mining the network value of customers. In *Proc. of KDD*, 2001.

[4] H. Ebel, L.I. Mielsch, and S. Bornholdt. Scale-free topology of email networks. *Phys. Rev.*, E 66:035103R, 2002.

[5] L. Garton, C. Haythornthwaite, and B. Wellman. Studying online social networks. *JCMC*, 3(1), 1997.

[6] S. A. Golder and B. A. Huberman. The structure of collaborative tagging systems. Technical report, HP Labs, 2005. http://www.hpl.hp.com/research/idl/papers/tags/.

[7] T. Hayward. Is digg being rigged: More data. http://taylorhayward.org/digggaming.html, September 2006.

[8] H. Kautz, B. Selman, and M. Shah. Referralweb: Combining social networks and collaborative filtering. *Communications of the ACM*, 4(3):63–65, 1997.

[9] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146, New York, NY, USA, 2003. ACM Press.

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

[11] K. Lerman. Social networks and social information filtering on digg. In *Proc. of International Conference on Weblogs and Social Media (ICWSM-07)*, 2007.

[12] K. Lerman and A. Galstyan. Mathematical model of foraging in a group of robots: Effect of interference. *Autonomous Robots*, 13(2):127–141, 2002.

[13] K. Lerman, Chris V. Jones, A. Galstyan, and Maja J. Matarić. Analysis of dynamic task allocation in multi-robot systems. *International Journal of Robotics Research*, 25(3):225–242, 2006.

[14] K. Lerman and Laurie Jones. Social browsing on flickr. In *Proc. of International Conference on Weblogs and Social Media (ICWSM-07)*, 2007.

[15] K. Lerman, A. Martinoli, and A. Galstyan. A review of probabilistic macroscopic models for swarm robotic systems. In Sahin E. and Spears W., editors, *Swarm Robotics Workshop: State-of-the-art Survey*, number 3342 in LNCS, pages 143–152. Springer-Verlag, Berlin Heidelberg, 2005.

[16] K. Lerman, A. Plangrasopchok, and C. Wong. Personalizing results of image search on flickr. In *AAAI workshop on Intelligent Techniques for Web Personlization*, 2007.

[17] J. Leskovec, L.A. Adamic, and B.A. Huberman. The dynamics of viral marketing. In *EC '06: Proceedings of the 7th ACM conference on Electronic commerce*, pages 228–237, New York, NY, USA, 2006. ACM Press.

[18] K. Maney. Wisdom of crowds. USA Today, September 12 2006. http://www.usatoday.com /tech/columnist/kevinmaney/2006-09-12-wisdom-of-crowdsx.htm.

[19] C. Marlow, M. Naaman, d. boyd, and M. Davis. Ht06, tagging paper, taxonomy, flickr, academic article, toread. In *Proceedings of Hypertext 2006*, New York, 2006. ACM, New York: ACM Press.

[20] A. Martinoli, K. Easton, and W. Agassounon. Modeling of swarm robotic systems: A case study in collaborative distributed manipulation. *Int. Journal of Robotics Research*, 23(4):415–436, 2004.

[21] P. Mika. Ontologies are us: A unified model of social networks and semantics. In *nternational Semantic Web Conference (ISWC-05)*, 2005.

[22] M.E.J. Newman. The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci. USA*, 98:404–409, 2001.

[23] E. Otte and R. Rousseau. Social network analysis: a powerful strategy, also for the information sciences. *Journal of Information Science*, 28(6):441–453, 2002.

[24] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[25] A. Papoulis. *Probability and Statistics*. Prentice Hall, Englewood Cliffs, NJ, 1990.

[26] S. Perugini, M. Andr Gonalves, and E. A. Fox. Recommender systems research: A connection-centric survey. *Journal of Intelligent Information Systems*, 23(2):107 – 143, September 2004.

[27] K. Rose. talk presented at the Web2.0 Conference, November 10 2006.

[28] K. Rose. Digg friends. http://diggtheblog. blogspot.com/2006/09/digg-friends.htm, September 2006.

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

[30] J. Warren and J. Jurgensen. The wizards of buzz. Wall Street Journal online, Feb 2007.

[31] F. Wu and B.A. Huberman. Novelty and collective attention. Technical report, Information Dynamics Laboratory, HP Labs, 2007.