Dynamics of Collaborative Document Rating Systems

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

ABSTRACT

The rise of social media sites, such as blogs, wikis, Digg and Flickr among others, underscores a transformation of the Web to a participatory medium in which users are actively creating, evaluating and distributing information. The social news aggregator Digg allows users to submit links to and vote on news stories. Like other social media sites, Digg also allows users to designate others as "friends" and easily track friends' activities: what new stories they submitted, commented on or liked. Each day Digg selects a handful of stories to feature on its front page. Rather than rely on the opinion of a few editors, Digg aggregates opinions of thousands of its users to decide which stories to promote to the front page. We construct a mathematical model to study how collaborative rating and promotion of news stories emerges from independent decisions made by many users. The model takes into account user behavior: *e.g.*, whether they read stories on the front page or through the Friends interface. Solutions of the model qualitatively reproduce the observed dynamics of votes received by actual stories on Digg.

Digg also ranks users according to how successful they are in getting their stories promoted to the front page. We create a model that describes how a user's rank changes in time as he gets more stories to the front page and becomes more influential in the community. We find qualitative agreement between predictions of the model and the evolution of rank for Digg users.

The Digg model of allowing users to collectively evaluate how interesting the news stories are can be generalized to collaborative evaluation of the quality of information. Mathematical analysis can be used as a tool to explore different voting methods to select the most effective one before the method is ever implemented in a real system.

Keywords

news aggregation, social networks, collaborative rating, mathematical analysis

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Joint 9th WEBKDD and 1st SNA-KDD Workshop '07 (WebKDD/SNA-KDD'07) August 12, 2007, San Jose, California, USA
Copyright 2007 ACM 978-1-59593-848-0 ...$5.00.

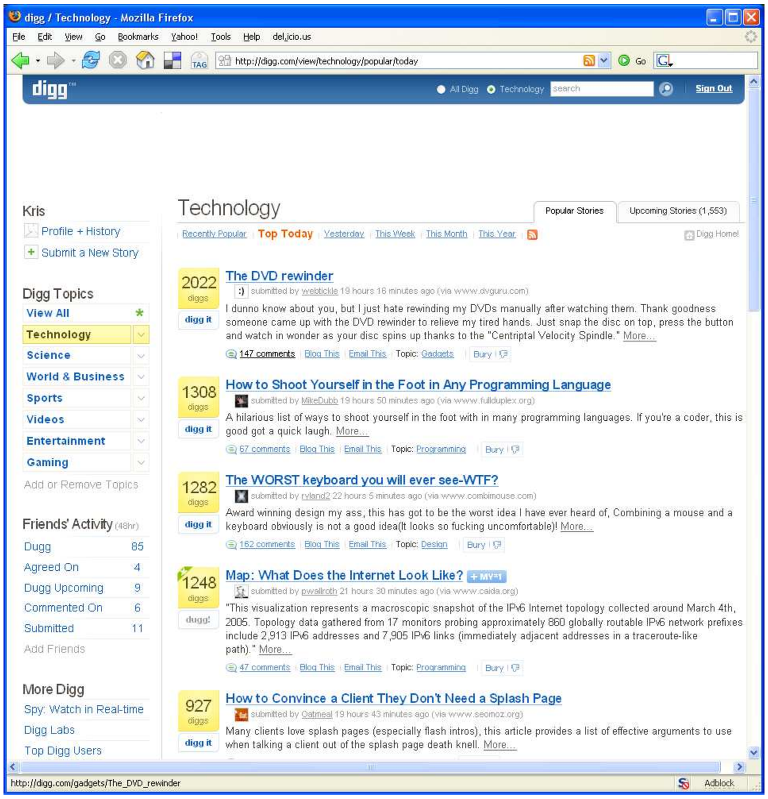
1. INTRODUCTION

The new social media sites — blogs, wikis, MySpace, Flickr, del.icio.us, and their ilk — have enjoyed phenomenal success in recent years. The extraordinary rise in their popularity is underscoring a transformation of the Web to a participatory medium where the users are actively creating, evaluating and distributing information. These sites share four elements: (1) users create or contribute content, (2) users annotate content with tags, (3) users evaluate content and (4) users create social networks by designating other users with similar interests as friends or contacts. These innovations help users solve hard information processing problems collaboratively, *e.g.*, detect public opinion trends in the blogosphere [1], construct taxonomies from the distributed tagging activities of many individuals [10], and use social networks as recommendation [5] and browsing aides [7].

One of the outstanding problems in information processing is how to evaluate the quality of documents or information in general. This problem crops up daily in information retrieval and Web search, where the goal is to find the information among the terabytes of data accessible online that is most relevant to a user's query. The standard practice of search engines is to identify all documents that contain user's search terms and rank results. Google revolutionized Web search by exploiting the link structure of the Web — created through independent activities of many Web page authors — to evaluate the contents of information on Web pages. Google, therefore, relies on an implicit rating scheme, where a link to a document is interpreted as a vote for it. Best seller lists are another example of an implicit rating system. An alternative to this is explicit rating, where a user assigns a rating, or a positive (or negative) vote to some document. Explicit ratings are used in many commercial collaborative filtering applications, on Amazon and Netflix for example, to recommend new products to users.

Social news aggregators like Digg[1] became popular in part because they rely on the distributed opinions of many independent voters to help users find the most interesting news stories. The functionality of Digg is very simple: users submit stories they find online, and other users rate them by voting on them. Each day Digg selects a handful of stories to feature on its front pages. Although the exact formula for how a story is selected for the front page is kept secret, so as to prevent users from "gaming the system" to promote fake stories or spam, it appears to take into account the number of votes a story receives. The promotion mechanism, therefore, does not depend on the decisions of a few editors,

[1] http://digg.com



Dugg	85
Agreed On	4
Dugg Upcoming	9
Commented On	6
Submitted	11

Add 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: Digg front page showing the technology section

but emerges from the activities of many users. This type of collective decision making can be extremely efficient, 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, 20 minutes before Google News showed it [12].

Designing a complex system like Digg, that exploits the emergent behavior of many independent evaluators, is exceedingly difficult. The choices made in the user interface, e.g., whether to allow users to see stories their friends voted on or the most popular stories within the past week or month, can have a dramatic impact on the behavior of the system and on user experience. Outside of running the system or perhaps simulating it, designers have little choice in how they evaluate the performance of different designs. Mathematical analysis can be used as a tool to explore the design space of collaborative rating algorithms to find parameters that optimize a given set of metrics (story timeliness vs interest, etc.), or eliminate unintended artifacts, before the algorithms are ever implemented in a real system.

This paper studies collaborative rating of news stories on Digg. Although Digg focuses on news stories and blogs, its collaborative rating approach can be extended to evaluating other kinds of information. We present a mathematical model of the dynamics of collective voting on Digg and show that solutions of the model strongly resemble votes received by actual news stories. By submitting and voting on stories, Digg users are also ranked by Digg. We present a second model that describes how a user's rank changes in time. We show that this model appears to explain the observed behavior of user rank.

The paper is organized as follows: In Section 2 we describe Digg's functionality and features in detail. In Section 3 we develop a model of collective voting. We compare solutions of the model to the behavior of actual stories. In Section 4 we develop a model of the dynamics of user rank and compare its solutions to the observed changes in user rank. In Section 5 we discuss limitations of mathematical modeling and identify new directions.

2. ANATOMY OF DIGG

Digg is a social news aggregator that relies on users to submit and moderate stories. 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. They are displayed in reverse chronological order of being submitted, 15 stories to the page, with the most recent story at the top. The story's title is a link to the source, while clicking on the number of diggs 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.

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 a 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 front page, therefore, emerges by consensus between many independent users.

Digg allows users to designate others as friends and makes it easy to track friends' activities.[2] The left column of the front page in Figure 1 summarizes the number of stories friends have submitted, commented on or liked (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 — *social navigation and filtering*. Rather than actively searching for new interesting content, or subscribing to a set of predefined topics, users can now put others to task of finding and filtering information for them.

Top users list.

Until February 2007 Digg ranked users according to how many of the stories the user submitted were promoted to the front page. Clicking on the Top Users link allowed one to browse through the ranked list of users. There is speculation that ranking increased competition, leading some users to be more active in order to improve their position on the Top users list. Digg discontinued making the list publicly available, citing concerns that marketers were paying top users to promote their products and services [14], although it is now available through a third party [3].

Social recommendation.

The Friends interface allows Digg users to see the stories their friends submitted or liked recently; therefore, it acts as a social recommendation system. By comparing users who voted on the story with the social network of the submitter, we showed that users tend to like (and vote on) the stories their friends submit and to a lesser extent, they tend to like the stories their friends like [5].

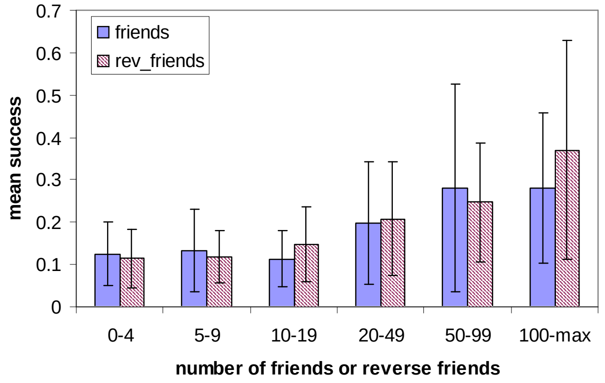


Figure 2: User's success rate vs size of their social network

Social networks on Digg contribute to how successful a user is at getting his stories promoted to the front page. A user's success rate is defined as the fraction of the stories the user has 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. We only include users who have submitted 50 or more stories in the analysis (total of 514 users). Users's mean success rate vs the size of their social network is shown in Figure 2. Data was binned to improve statistics. There is a significant correlation between users's success rate and the number of reverse friends they have.

3. COLLABORATIVE RATING DYNAMICS

In order to study how the front page emerges from independent opinions of many users, we tracked both the upcoming and front page stories in Digg's technology section. We collected data by scraping Digg site with the help of Web wrappers, created using tools provided by Fetch Technologies[4]:

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.

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 the user has submitted, commented and voted on; number of stories promoted to the front page; users'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 in May and in July 2006.

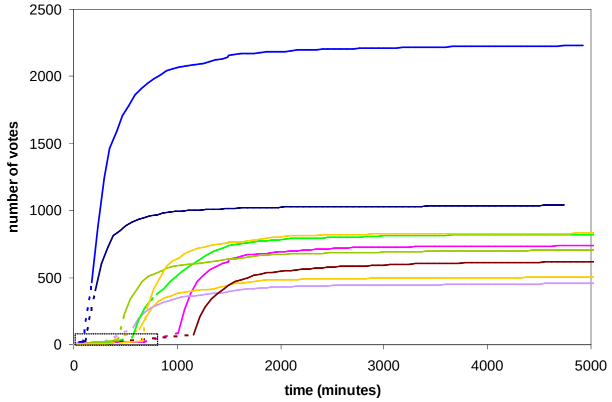


Figure 3: Dynamics of votes of select stories over a period of four days. The small rectangle in the lower corner highlights votes received by stories while in the upcoming stories queue. Dashes indicate story's transition to the front page.

We identified stories that were submitted to Digg over the course of approximately one day and followed them over a period of several days. Of the 2858 stories submitted by

[2] Note that the friend relationship is asymmetric. When user A lists user B as a *friend*, user A is able to watch the activity of B but not vice versa. We call A the *reverse friend* of B.

[3] http://www.efinke.com/digg/topusers.html

[4] http://fetch.com/

1570 distinct users, only 98 stories by 60 users made it to the front page. Figure 3 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, and once promoted to the front page, it accumulates votes at a much faster rate. As the story ages, accumulation of new votes slows down, and the story's rating saturates at some value. This value depends on how *interesting* the story is to the Digg community.

It is worth noting that the top-ranked users are not submitting the most interesting stories (that get the most votes). Slightly more than half of the stories our data set came from 14 top-ranked users (rank< 25) and 48 stories came 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. A second observation is that top-ranked users are responsible for multiple front page stories. A look at the statistics about top users provided by Digg shows that this is generally the case: of the more than 15,000 front page stories submitted by the top 1020 users, the top 3% of the users are responsible for 35% of the stories.

3.1 Mathematical model

In this section we present a mathematical model that describes the evolution of the number of votes received by a story. Our goal is not only to produce a model that can explain — and predict — the dynamics of collective voting on Digg, but one that can also be used as a tool to study the emergent behavior of 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. This is simply an *ad hoc* parameter to characterize how relevant or interesting a story is to Digg audience. The number of votes a story receives depends on its *visibility*, which simply means how many people can see a story and follow the link to it. The factors that contribute to the story's visibility include:

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

Digg offers additional ways to see popular stories: e.g., most popular stories submitted over the preceding week or month. We assume that these browsing modalities do not generate significant views, and focus on the simpler model that takes into account solely the factors enumerated above.

3.1.1 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 visit Digg and proceed to page 2, 3 and so on, 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 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 farther 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. The change in a story's current page number in the May data set can be fit by lines $\{p, q\} = k_{\{u,f\}} t$ with slopes $k_u = 0.060$ pages/m (3.60 pages/hr) for the upcoming stories and $k_f = 0.003$ pages/m (0.18 pages/hr) for the front page stories.

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 in the upcoming queue; when $m \geq h$, the story is visible on the front page. This seems to approximate Digg's promotion algorithm as of May 2006: in our data set we did not see any front page stories with fewer than 44 votes, nor upcoming stories with more than 42 votes.

3.1.2 Visibility through the Friends interface

The Friends interface offers the user ability to see the stories his friends have (i) submitted, (ii) liked (voted on), (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 features closely approximate the functionality offered by other social media sites: for example, 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 the story's submitter has. These are the users who are watching the submitter's activities. We assume that these users visit Digg daily, and since they are likely to be geographically distributed across many time zones, they see the new story at an hourly rate of $a = S/24$. The story's visibility through the submitter's social network is therefore $v_s = a\Theta(S - at)\Theta(48 - t)$. $\Theta(x)$ 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 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 voted on, it becomes visible to more users through the see the "stories my friends dugg" part of the Friends interface. Although the empirical value of S_m, the combined social network of the first m voters, is highly variable from story to story in our data set, it's average value (averaged over 195 stories) has consistent growth: $S_m = 112.0 * log(m) + 47.0$. The story's visibility through the friends of voters is, therefore, $v_m = bS_m\Theta(h - m)\Theta(48hrs - t)$, where b is a scaling factor that depends on the length of the time interval: for hourly counts, it is $b = 1/24$.

3.1.3 Dynamics of ratings

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

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

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; otherwise, it 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 only, 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 stories queue q and the front page p change in time according to:

$$
\begin{aligned}
p(t) &= (k_f(t-T_h)+1)\Theta(T_h-t) &(5)\\
q(t) &= k_u t + 1 &(6)
\end{aligned}
$$

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

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

3.2 Solutions

We solve Equation 7 subject to the initial conditions $m(t=0)=1$, $q(t=0)=1$, as 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 change from one submission to another. Therefore, we fix values of the parameters $c=0.3$, $c_u=0.3$ and $c_f=0.3$ and study the effect r and S have on the number of votes the 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 $h=40$.

Our first observation is that introducing social recommendation via the Friends interface allows stories with smaller r to be promoted to the front page. In fact, we can obtain an analytic solution for the maximum number of votes a story can receive on the upcoming stories queue without the social filtering effect being present. We set $v_f=v_s=v_m=0$ and convert Equation 7 to a differential form by taking $\Delta t \to 0$:

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

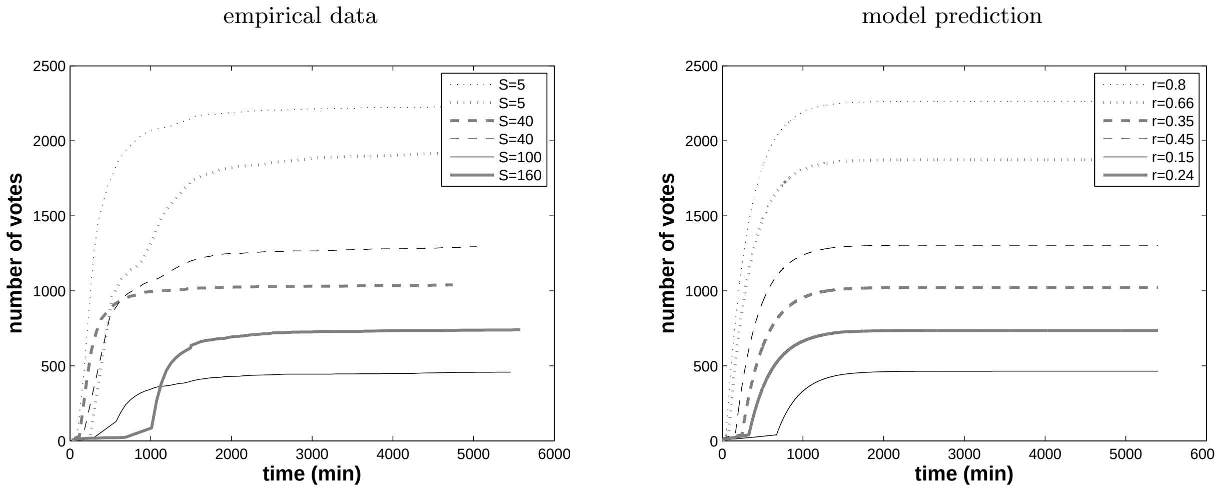
The solution of the above equation is $m(T) = rcN(c_u^{k_u T}-1)/(k_u \log c_u)+1$. Since $c_u<1$, the exponential term will vanish for large times and leave us with $m(T\to\infty) = -rcN/(k_u\log c_u)+1 \approx 42r+1$. Hence, the maximum rating a story can receive on the upcoming pages only is 43. Since 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. On average, the more reverse friends the submitter has, the smaller the minimum interestingness required for a story he submits to be promoted to the front page. Conversely, users with few reverse friends will generally have only the very interesting stories promoted to the front page. The second observation is that the more interesting story is promoted faster than a less interesting story.

Next, we consider the second modality of the Friends interface which allows users to see the stories their friends voted on. This is the situation described by the model Equation 7. Figure 4(a) shows the evolution of the number of votes received by six real stories from our data set. S denotes the number of reverse friends the story's submitter had at the time of submission. Figure 4(b) shows solutions to Equation 7 for the same values of S and values of r chosen to produce the best fit with the data. 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 bottom two lines, for stories submitted by users with $S=100$ and $S=160$. The difference between data and the model is not surprising, given the number of approximations made in the course of constructing the model (see Section 5). For example, we assumed that the combined social network of voters grows at the same rate for all stories. This obviously cannot be true. If the combined social network grew at a slower than assumed rate for the story posted by the user with $S=160$, then this would explain the delay in being promoted to the front page. Another effect not currently considered is that a story may have a different interestingness value for users within the submitter's social network than to the general Digg audience. The model can be extended to include inhomogeneous r.

3.3 Modeling as a design tool

Designing a collaborative rating system like Digg, which exploits the emergent behavior of many independent evaluators, is exceedingly difficult. The choices made in the user interface can have a dramatic impact on the behavior of the collaborative rating system and on user experience. For example, should the users see the stories their friends voted on or the week's or the month's most popular stories? The designer has to consider also the tradeoffs between story timeliness and interestingness, how often stories are promoted. The promotion algorithm itself can have a dramatic impact on the behavior of the collaborative rating system. Digg's promotion algorithm (prior to November 2006) alienated some users by making them feel that a cabal of top users controlled the front page. Changes to the promotion algorithm appear to have alleviated some of these concerns (while perhaps creating new ones). Unfortunately, there are few tools, short of running the system, that allow developers to explore the various choices of the promotion algorithm.

We believe that mathematical modeling and analysis can be a valuable tool for exploring the design space of collaborative rating systems, despite the limitations described in Section 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, *e.g.*, by setting the promotion



(a)

(b)

Figure 4: (a) Evolution of the number of votes received by six stories from the May data set. The number of reverse friends the story submitter has is given by S. (b) Predictions the model makes for the same values of S as (a).

threshold to be a function of the number of reverse friends the submitter has.

4. DYNAMICS OF USER RANK

From its inception until February 2007, Digg ranked users according to how successful they were in getting their stories promoted to the front page. The more front page stories a user had, the higher was his standing ($rank = 1$ being the highest). If two users had an equal number of front page stories, the one who was more active (commented and voted on more stories) had higher rank. The Top Users list was publicly available and offered prestige to those who made it into the top tier. In fact, it is widely believed that improving ones rank, or standing within the community, motivated many Digg users to devote significant portions of their time to submitting, commenting on and reading stories. Top users garnered recognition as other users combed the Top Users list and made them friends. They came to be seen as influential trend setters whose opinions and votes were very valuable [14]. In fact, top users became a target of marketers, who tried to pay them to promote their products and services on Digg by submitting or voting on content created by marketers. In an attempt to thwart this practice, in February 2007 Digg discontinued making the Top Users list publicly available.

We are interested in studying the dynamics of user rank within the Digg community. For our study we collected data about the top 1,000 ranked Digg users weekly from May 2006 to February 2007. For each user we extracted user's rank, the number of stories the user submitted, commented and voted on, the number of stories that were promoted to the front page, and the number of user's friends and reverse friends ("people who have befriended the user"). Figure 5 shows the change in rank of six different users from the data set. The top ranked user ($user2$) managed to hold on to that position for most of time, but $user6$, who was ranked second at the beginning of the observation period saw his rank slip to 10. Some users, such as $user1$ and $user5$, came in with low rank but managed to reach the top tier of users by week 20. Others ($user4$ and $user6$) saw their rank stagnate.

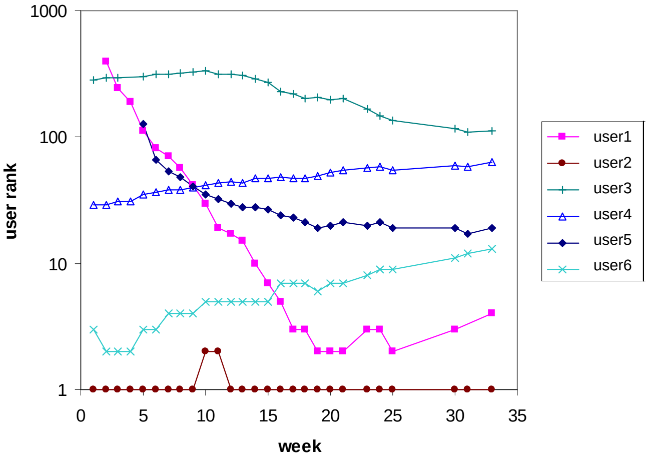
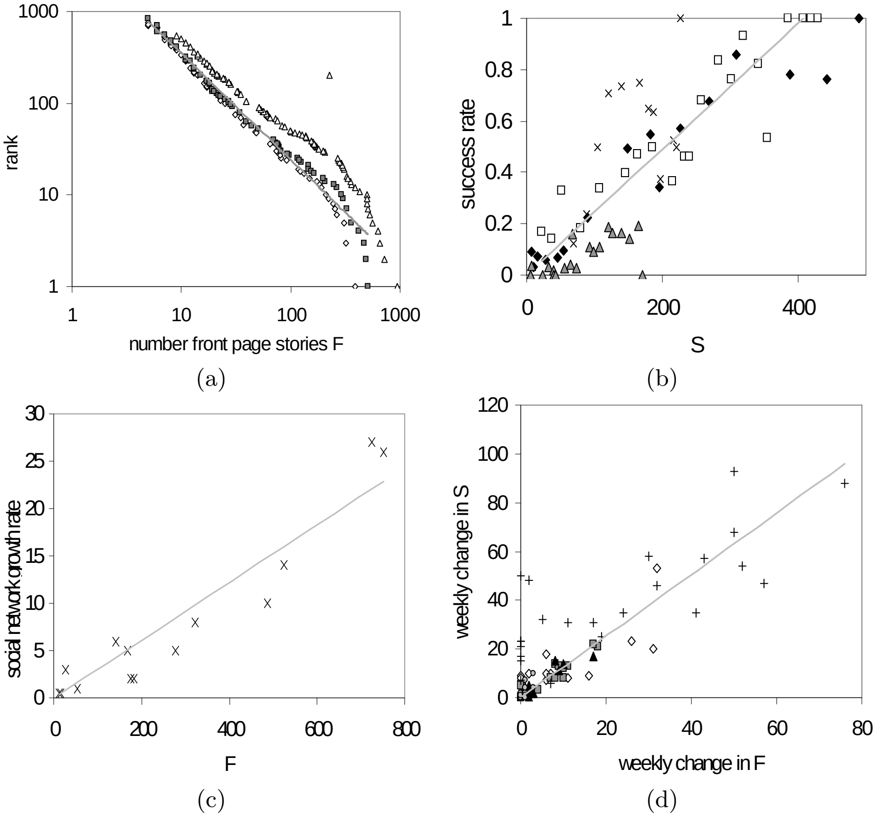


Figure 5: Evolution of user rank

4.1 Mathematical model

We are interested in creating a model that can predict how a user's rank will change in time based on the user's activity level. The model also describes the evolutions of the user's personal social network, or the number of reverse friends. In addition to its explanatory power, the model can be used to detect anomalies, for example, cases when a user's rank, or social network, changes faster than expected due to collusion with other users or other attempts to game the community. Because we do not know the exact formula Digg uses to compute rank, we will use F, the number of user's front page stories, as a proxy for rank. Figure 6(a) plots user's rank vs the number of front page stories for three randomly chosen users. The data is explained well by a power law with exponent -1: i.e., $rank \propto 1/F$.



(a) (b) (c) (d)

Figure 6: Parameter estimation from data. (a) Users' rank vs number of their stories that have been promoted to the front page. (b)Different users' success rates at getting their stories promoted to the front page vs the number of reverse friends they have. In all plot, solid lines represent fit to the data. (c) Temporal growth rate of the number of user's reverse friends as a function of user rank for the weeks when no new stories were submitted by these users. (d) Weekly change in the size of the social network vs newly promoted front page stories.

The number of stories promoted to the front page clearly depends on the number of stories a user submits, with the proportionality factor based on the user's success rate. A user's success rate is simply the fraction of the newly submitted stories that are promoted to the front page. As we showed above, a user's success rate is linearly correlated with the number of reverse friends he has — what we call social network size S. If M is the rate of new submissions made over a period of time Δt =week, then the change in the number of new front page stories is

$$\Delta F(t) = cS(t)M\Delta t \tag{9}$$

To estimate c, we plot user's success rate vs S for several different users, as shown in Figure 6(b). Although there is scatter, a line with slope $c = 0.002$ appears to explain most of the trend in the data.

A given user's social network S is itself a dynamic variable, whose growth depends on the rate other users discover him and add him as a friend. The two major factors that influence a user's visibility and hence growth of his social network are (i) his new submissions that are promoted to the front page and (ii) his position on the Top Users list. In addition, a user is visible through the stories he submits to the upcoming stories queue and through the comments he makes. We believe that these effects play a secondary role to the two mentioned above. The change in the size of a user's social network can be expressed mathematically in the following form:

$$\Delta S(t) = g(F)\Delta t + b\Delta F(t) \tag{10}$$

In order to measure $g(F)$, how a user's rank affects the growth of his social network, we identified weeks during which some users made no new submissions, and therefore, had no new stories appear on the front page. In all cases, however, these users' social networks continued to grow. Figure 6(c) plots the weekly growth rate of S vs F. There is an upward trend indicating that the higher the user's rank (larger F) the faster his network grows. The grey line in Figure 6(c) is a linear fit to the data of functional form $g(F) = aF$ with $a = 0.03$. Figure 10(d) shows how newly promoted stories affect the growth in the number of reverse friends for several users. Although there is variance, we take $b = 1.0$ from the linear fit to the data.

4.2 Solutions

Figure 7 shows how the personal social network (number of reverse friends) and the number of front page stories submitted by six different users from our data set change in time. The users are the same ones whose rank is shown in Figure 5. To the right of each graph we plot solutions to Equation 9 and Equation 10. The equations were solved

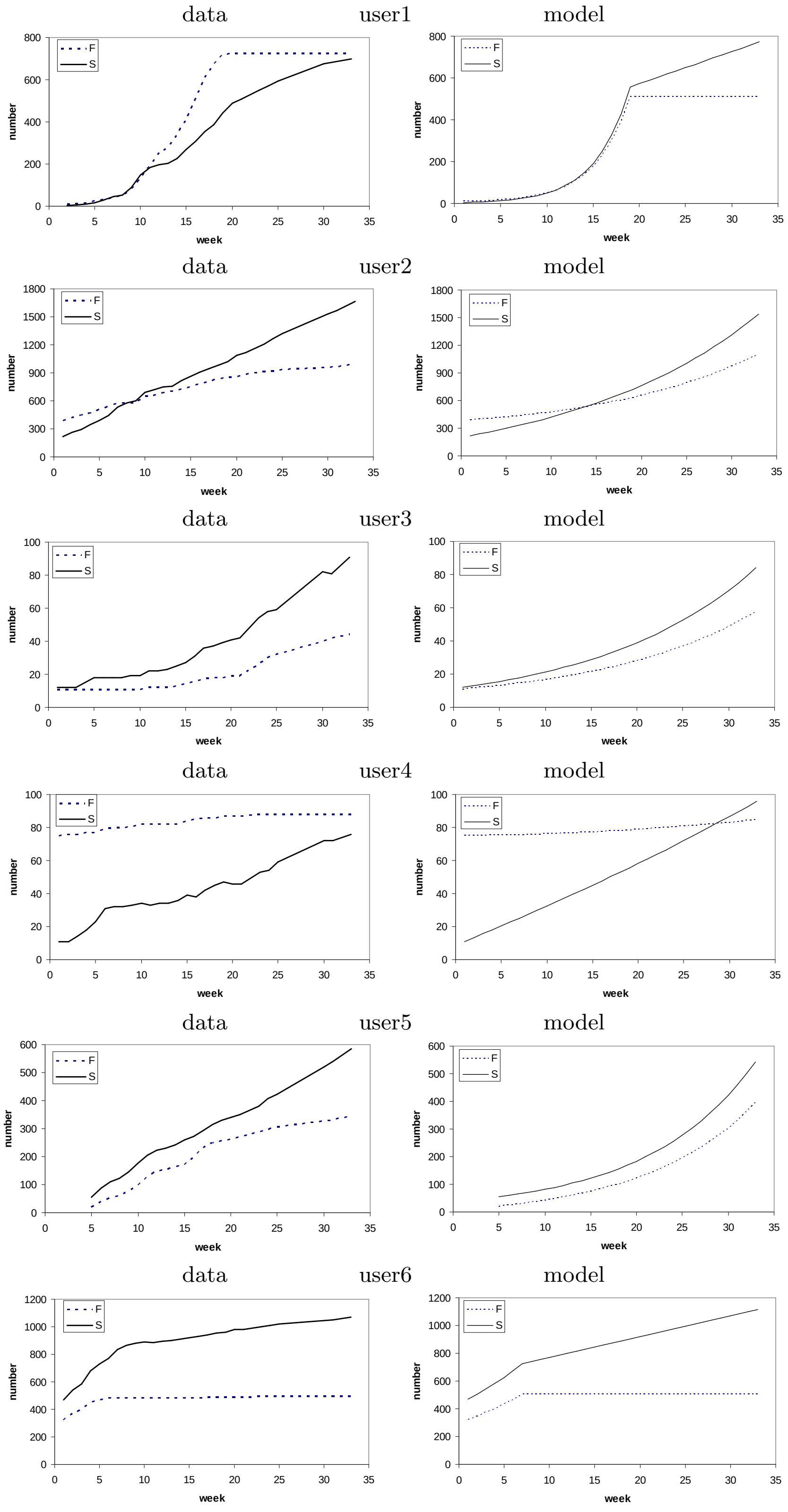


Figure 7: Change over the course of 25 weeks of the number of front page stories and the number of reverse friends a user has. The six users are the same ones shown in Figure 5. The right hand plots show solutions to the rank dynamics model for each user.

under the initial conditions that F and S take the values they have at the beginning of the tracking period for that user. We kept the submission rate M fixed at its average weekly value over the tracking period. The actual submission rate fluctuates significantly for a given user from week to week. We expect that including the actual submission rate will substantially improve the agreement between the model and the data.

Solutions to the model qualitatively reproduce the important features of the evolution of user's rank and social network. Two factors — user's activity via new submissions and the size of his social network — appear to explain the change in user's rank. As long as the user stays active and contributes stories to Digg, as exemplified by *user*2, *user*3, *user*4 and *user*5, both the number of promoted stories (rank) and the size of the user's social network continue to grow. If a user stops contributing stories, *user*1 and *user*6, his rank will stagnate as F remains constant, while his social network continues to grow, albeit at a slower rate. Although a user can choose to submit more or fewer stories to Digg, he cannot control the growth of his social network, e.g ., how and when other users choose to make him a friend.[5] This helps promote independence of opinions, a key requirement of the collaborative rating process, and raise the quality of ratings. It appears, however, that the Top Users list serves to cement the top tier position of the highest ranked users, since they continue to grow their social networks, which in turn improves their success rate. It will be interesting to observe how elimination of the Top Users list alters the Digg community and the quality of stories that appear on the front page.

5. LIMITATIONS

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. For example, 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 voted on. There are other browsing modalities on Digg that we did not include 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, in the Friends interface, a user can also see the stories his friends have commented on or that are still in the upcoming queue, as well as the stories they have submitted or voted on. We chose to include only the latter two options from the Friends interface in our model.

In addition to the model structure, we made a number of assumptions about the form of the terms and the parameters. The first model describes the dynamics of votes an *average* story receives. In other words, it does not describe how the rating of a specific story changes in time, but the votes on many similar stories averaged together. Another point to keep in mind is that 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 the parameter N in the model. We also assume that a story's interestingness is the same for all users. In the model for rank dynamics, all parameters were characterized by single value — taken to be the mean or characteristic value of the distribution of user behavior. In future work we intend 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 models 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 models we present in the paper do include the most salient features of the Digg users' behavior. We showed that the models qualitatively explain some features of 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. We plan to address these issues in future research.

6. PREVIOUS RESEARCH

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 [4] 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 [3, 11]. 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 information.

Social navigation, a concept closely linked to CF, helps users evaluate the quality of information by exposing information about the choices made by other users "through information traces left by previous users for current users" [2]. Exposing information about the choices made by others has been has been shown [13] 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 only global information about the preferences of others (similarly to the best seller lists and Top Ten albums). We believe that exposing local information about the choices of others within your community can lead to more effective collective decision making.

Wu and Huberman [15] have recently studied the dynamics of collective attention on Digg. They proposed a simple

[5] We suspect that a user is able to influence the growth of his social network through the implicit social etiquette of reciprocating friend requests, but we have not yet been able to prove this conjecture.

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 diggs 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 the story spends on the first front page. The problem studied by Wu and Huberman is complementary to ours, as they studied dynamics of stories *after* they hit the front page. The authors did not identify a mechanism for the spread of interest. We, on the other hand, propose social networks as a mechanism for spreading stories' visibility and model evolution of diggs both before and after the stories hit 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 [8]. This framework was successfully applied to study collective behavior in groups of robots [9, 6]. 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.

7. 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 which can then be used to get new recommendations for content or documents. Another novel feature is the collaborative evaluation of content, either explicitly through voting or implicitly through user activity.

We studied collaborative ranking of content on Digg, a social news aggregator. 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. We also studied how user's rank, which measures the influence of the user within the community, changes in time as the user submits new stories and grows his social network. Again we found qualitative agreement between data and model predictions.

Besides offering 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. It can help designers investigate global consequences of different story promotion algorithms before they are implemented. Should the promotion algorithm depend on a constant threshold or should the threshold be different for every story? Should it take into account the story's timeliness or the popularity of the submitter, etc.?

Acknowledgements

This research is based on work supported in part by the National Science Foundation under Award Nos. IIS-0535182 and IIS-0413321. We are grateful to Fetch Technologies for providing wrapper building and execution tools.

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

[4] 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.

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

[6] 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.

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

[8] 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.

[9] 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.

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

[11] 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.

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

[13] 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.

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

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