

Analysis of Social Voting Patterns on Digg

Kristina Lerman and Aram Galstyan
University of Southern California
Information Sciences Institute
4676 Admiralty Way
Marina del Rey, California 90292
{lerman,galstyan}@isi.edu

ABSTRACT

The social Web is transforming the way information is created and distributed. Authoring tools, e.g., blog publishing services, enable users to quickly and easily publish content, while social networking sites, such as Digg, Del.icio.us, YouTube, and Facebook, may be used to distribute content to a wider audience. With content fast becoming a commodity, interest in using social networks to promote content has grown, both on the side of content producers (viral marketing) and consumers (recommendation). We study the role of social networks in promoting content on the Digg. Digg is a social news aggregator that allows users to submit links to and vote on news stories. Digg, therefore, provides a direct measure of content quality, specifically, how interesting a news story is is judged by the number of votes it receives. Digg's goal is to feature the most interesting stories on its front page. Like other social networking sites, Digg allows users to designate others as "friends" and easily track friends' activities: what new stories they submitted or voted on recently. Digg is unique in that it makes this information, along with the votes, publicly available. We empirically studied the spread of interest in news stories submitted to Digg in June 2006. We found that not only do social networks play a significant role in promoting news stories, but that we can also predict how many votes a story will receive over a long period of time by monitoring the spread of early interest in the story.

1. INTRODUCTION

The social Web, a label that includes both the social networking sites MySpace and Facebook and the social media sites such as Digg, Flickr, is changing the way content is created and distributed. Web-based authoring tools enable users to rapidly publish content, from stories and opinion pieces on weblogs, to photographs and videos on Flickr and YouTube, to advice on Yahoo Answers, to Web discoveries on Del.icio.us and Furl. Such user-generated content accounts for much of the new Web content. In addition to allowing users to share content, social Web sites also include a

social networking component, which means that they allow users to mark other users as friends or contacts, and provide an interface to track their friends' activities, e.g., the new content they created.

With the commodification of content, content producers face a dilemma, namely how to most effectively promote and distribute their content. The question facing content consumers is how to efficiently identify interesting or relevant content in a vast stream of user-generated content. Marketers and social scientists have long recognized the central role social networks play in the spread of information [8]. Before the advent of electronic communication, 'word-of-mouth' recommendation was carried out mainly through telephone and letter communications, or personal interactions. Modern communications technologies have further emphasized the role of social networks in information dissemination [24, 9] and product recommendation [4]. Researchers have also noted the similarity between these processes and the spread of epidemics on networks. This formulation allows researchers to draw on an extensive literature studying statistical properties of the dynamics of epidemics on networks [16, 15]. Despite its importance to social recommendation and viral marketing, there are few empirical studies of information propagation in social networks. Even more interestingly, existing studies have produced conflicting results. One study of music recommendation conducted in a laboratory setting, found that users' choice of music to listen to was significantly influenced by choices made by their peers [22]. However, a large-scale study of viral marketing on Amazon [13] showed word of mouth recommendations to be largely ineffective in leading to new purchases of products. Like the previous study of information propagation through email [24], it found that most recommendation chains terminate after just a few steps. The study did note the sensitivity of recommendation to price and category of product, leaving open the question of whether social networks are an effective tool for disseminating information about, and helping users discriminate between, free (or similarly priced) products. For example, is social recommendation effective in promoting product (e.g., movie, audiobook) in a subscription-based service like Netflix or Audible? Can does it affect the choices user makes about what (free) content to read online, such as news stories and blogs?

We study the role of social networks in information propagation on the social news aggregator Digg¹. Digg became

¹<http://digg.com>

popular partly because, rather than using an editorial board to find the most interesting stories online, it relies on the opinions of its multitude of users to identify the most newsworthy stories. Like other social Web sites, Digg allows its users to create social networks by designating other users as friends, and makes it easy to track friends' activities. Through the Friends Interface, a user can see the stories his friends submitted or liked recently. The Friends Interface, therefore, acts as a social recommendation engine, suggesting content to a user that his friends have found interesting.

In a previous study one of the authors used statistics about the activities of the top users to show that users with bigger social networks were more successful at getting their stories promoted to the front page [11]. This finding could be explained by social recommendation. Here we perform further empirical studies of social recommendation on Digg, by examining the impact of the social ties on the voting dynamics. Our results suggest that if in the initial stages of voting most of the votes come from within the social neighborhood of the story's submitter, then the total number of votes received by that story tends to be relatively small. In other words, these are stories which are propagated mostly through the network effect, but do not carry sufficient interest for the users outside the submitter's community. On the other hand, stories for which initial votes come from distant users, tend to generate much larger number of final votes. Thus, one can make predictions about the potential audience of a story simply by analyzing where the initial votes came from. Indeed, we build a simple regression tree classifier that takes only two inputs, the number of incoming links of the submitter and the number of in-network votes among the first ten, and show that it provides a reasonably good prediction accuracy.

2. DIGG'S FUNCTIONALITY

The social news aggregator Digg relies on users to submit and moderate news stories. Each new story goes to the upcoming stories queue. The new submissions (there are 1-2 new submissions every minute) are displayed in reverse chronological order, 15 to the page, with the most recent story at the top. Each day Digg selects a handful of stories to feature on its front page. Digg's goal is to identify and promote only the most interesting of the submitted stories, and it relies on its users opinions to find these stories. Digg's automatic story promotion algorithm looks at the voting patterns made within 24 hours of the stories' submission. Although its details are kept secret and change on a regular basis [20], the promotion algorithm seems to take into account the number of votes a story receives and the rate at which it receives them, among other factors. In the data we collected, we did not see any front page stories with fewer than 43 votes, nor did we see any stories in the upcoming queue with more than 42 votes.

Digg allows users to designate others as friends and makes it easy to track friends' activities. The friendship 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 *fan* of B . Digg provides a Friends Interface, which summarizes a user's friends' recent activity: the number of stories his friends have submitted, commented on or voted on in the preceding 48 hours. Tracking activities of friends is

a feature of many social Web sites and is one of their major draws.

Digg users vary widely in their activity levels. Some users casually browse the front page, voting on one or two stories. Others spend hours a day combing the Web for new stories to submit, and voting on stories they found on Digg. Digg calculated a users' reputation based on how successful they were in getting their stories promoted to the front page. In order to encourage activity, Digg publicized users' reputation on the *Top Users* list. A look at the statistics of user activity showed that top-ranked users were disproportionately active: of the more than 15,000 front page stories submitted by the top 1000 Digg users as of June 2006, the top 3% of the users were responsible for 35% of the submissions and a similarly high fraction of the votes cast and comments made. Digg discontinued ranking users by their reputation in February 2007 [21], although it is not clear whether this had an impact on users' activity [12].

2.1 Digg dataset

For our study, we scraped Digg's Technology section with the aid of a tool provided by Fetch Technologies. On June 30, 2006, we scraped Digg's front page, collecting data about roughly 200 of the most recently promoted stories. For each story, we extracted the story's title, name of the submitter, time the story was submitted, as well as names of the first 215 users to vote on the story. Although we do not have the time stamp of each vote, they are listed in chronological order, with submitter's name appearing first on the list. In February 2008 we augmented this data with the final vote count (number of diggs) the stories received. In all, we collected information about votes from over 16,600 distinct users.

The basic dynamics of votes received by stories appears the same [11]. While in the upcoming queue, a story accumulates votes at some slow rate, and once it is promoted to the front page, it accumulates votes at a much faster rate. As the story ages, the accumulation of new votes slows down, and after a few days, the story's vote count saturates at some value. This value depends on how *interesting* the story is to the general Digg community. Figure 1(a) shows the histogram of the final vote counts received by stories. Twenty percent of the stories were not very interesting, receiving fewer than about 500 votes, and twenty percent were very interesting, receiving more than 1500 votes. This graph is very similar to one presented in [25], which analyzed votes received by almost 30,000 front page stories on Digg submitted over a period of a year. That work suggested that it is most common for a story to receive between 400-600 votes.

The distribution of user activity is similarly skewed, as shown in Figure 1(b). While most of the users had one story promoted to the front page, a number of users were responsible for multiple submissions. These were also the users with highest reputation, the so-called top users. Voting statistics are even more skewed. While most of the users voted on only one story, some voted on many, and a few on well over a hundred stories.

2.2 Social networks

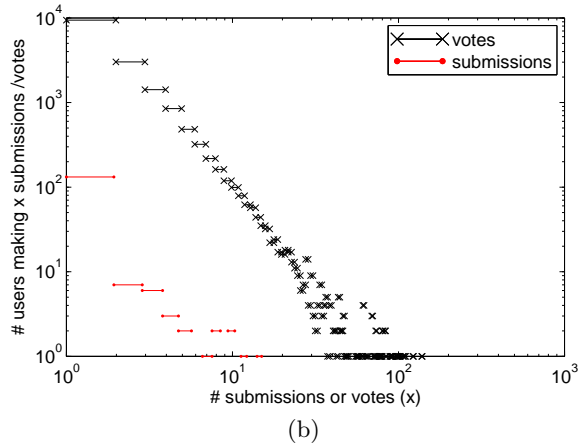
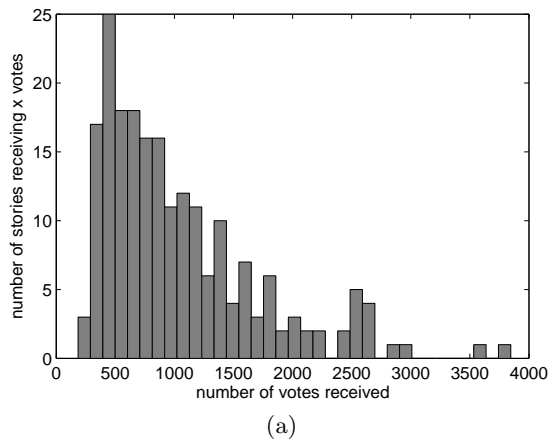


Figure 1: Statistics of story and user activity: (a) Histogram of the number of votes received by stories. (b) Histogram of the number of stories submitted and voted on.

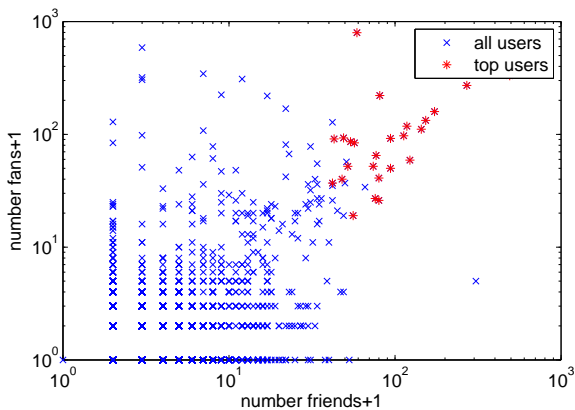


Figure 2: Scatter plot of the number of friends and fans of the voters in our dataset.

In addition to data about stories, we also extracted a snapshot of the social network of the top-ranked 1020 Digg users as of June 30, 2006. This data contained the names of each user’s friends and fans. As a reminder, user A ’s friends are all the users that A is watching (outgoing links on the social network graph), while A ’s fans are all the users watching his activity (incoming links). Since the original social network did not contain information about all the voters in our dataset, we augmented this data in February 2008 by extracting names of fans of the 15,000+ additional users. Many of these users acquired new fans between June 2006 and February 2008. Although Digg does not provide information about the time a fan link was created, it does list these links in reverse chronological order, with the most recent appearing on top. In addition to a fan’s name, Digg also gives the date the fan joined Digg. But eliminating fans who joined Digg after June 30, 2006, we believe we were able to faithfully reconstruct the fan links (incoming edges) for all the users in our dataset. Figure 2 shows the scatter plot of the number of friends vs fans each user has. The number

was offset by one in order to be plotted on the log-log plot. The red stars correspond to 25 highest ranked users in our dataset. As can be seen from this plot, these users have greater numbers of friends and fans than other users.

3. INFORMATION SPREAD IN NETWORKS

The Digg dataset allows us to empirically study the spread of information on social networks. Before a story reaches the front page, it is visible only on the upcoming stories queue and through the Friends interface. Although some users browse the upcoming stories queue, the quantity of submissions there (more than 1500 daily at the time we collected data) makes browsing unmanageable to most users. Digg also offers a visual interface to browse the upcoming and front page stories, Swarm and Stack. These visualizations are supposed to make it easier for users to identify more popular stories, but it is not clear how many users take advantage of them. Increasingly, many news sites and blogs are including a “Digg it” button to allow its readers to submit or vote on the story directly from the story’s Web page. Again, it is not clear how many users take advantage of this option. We believe that social networks play an important role in promoting stories on Digg. In a previous work we presented data to support the claim that users employ the Friends interface to filter the vast stream of new submissions to see the stories their friends liked. Below we study the information spread in detail.

3.1 Information cascades

At the time of submission, the story is visible only to submitter’s fans through the ‘see the stories your friends submitted’ part of the Friends interface. As the story receives new votes, it becomes visible to many more users through the ‘see the stories my friends dugg’ (voted on) part of the Friends interface. A story’s *influence* is given by the number of users who can see it through the Friends interface. Figure 3(a) shows a histogram of the story’s influence. Slightly more than half of the stories in our sample were submitted by poorly connected users with fewer than ten fans. After stories received ten new votes, almost half of them were visible to at least 200 users through the Friends interface. After

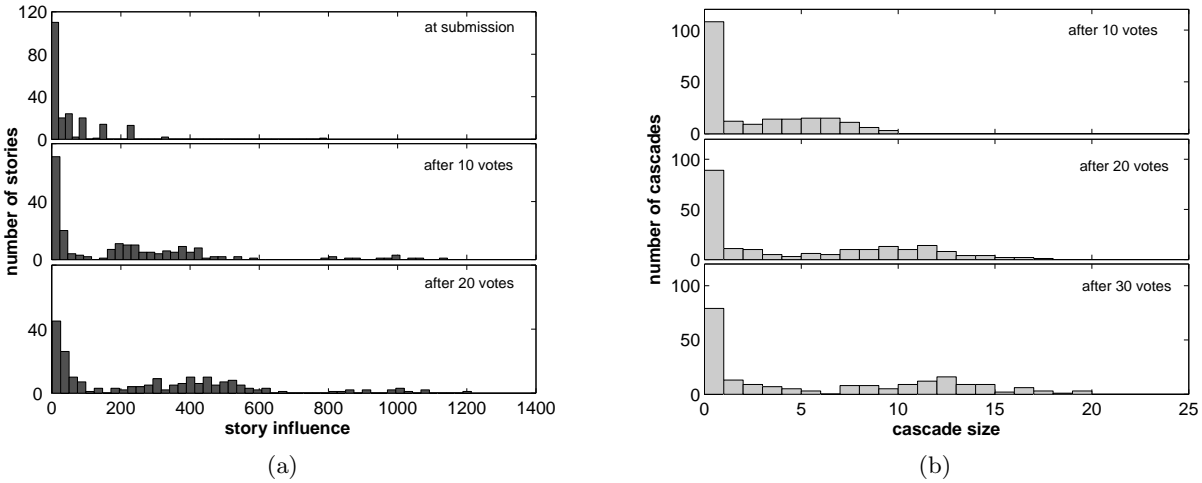


Figure 3: Spread of interest in stories: (a) histogram of the story’s influence, defined as the number of users who can see it through the friends interface, after it received ten votes, and (b) the number of in-network votes the story received within the first ten votes.

30 votes, all the stories in our sample were visible to at least ten other users through the Friends interface, and majority of the stories were visible to hundreds of users.

Because we know the social network of Digg users, we can compute how many votes came from within the network — from fans of the previous voters. This is the story’s *cascade*. Figure 3(b) shows the distribution of cascades in our sample. For 30% of the stories, at least half of the first 10 votes were in-network votes. Cascades grow with the number of votes cast. After 20 votes, 28% of the stories had at least 10 votes from fans of the previous voters’ and after 30 votes, 36% of the stories had at least 10 in-network votes.

4. STORY INTERESTINGNESS

The total number of votes a story receives gives a measure of how *interesting* it is to Digg’s audience. Based on a number of features, such as the number of votes received and the rate at which it receives them, Digg attempts to predict, within the first 40 or so votes, whether the story will be found interesting by its audience. It is especially challenging to Digg to predict how interesting a story submitted by one of the top users is. Top users are far more active and well connected than other users, meaning that they submit and vote on many more stories, some of which happen to be stories submitted by their friends. Since top users are more likely to be in the same network, their stories are more likely to get more votes and therefore, be promoted to the front page. In September 2006, a controversy about top user dominance [1, 2] caused Digg to modify the promotion algorithm to take into account “unique digging diversity of the individuals digging the story” [20]. Although this modification did result in changes in front page composition, it is not clear whether it affected the spread of interest in stories on the social networks on Digg. Rather than discounting the votes coming from fans, as Digg has chosen to do, we show that we can predict how interesting a story will be by monitoring its spread through the social network.

4.1 Social networks and interestingness

In a previous work [11] we showed that top Digg users were very successful in getting their stories promoted to the front page. We claimed that this could be explained by social browsing, i.e., the fact that Digg users use the Friends interface to find new interesting stories. We showed that social browsing, together with the observation that top users have more fans than other users, explains how less interesting stories submitted by top users are promoted to the front page. Here we study in detail how the spread of interest in a story through the social network relates to its interestingness.

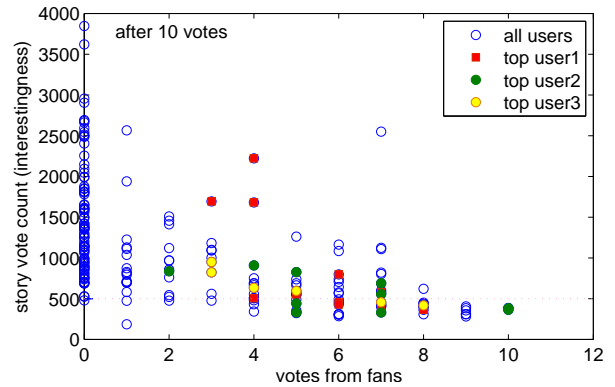


Figure 4: Number of votes stories receive from within the social network within the first ten votes vs the stories’ interestingness. Colored symbols correspond to stories submitted by select top users.

Figure 4 shows the total number of votes a story receives, its interestingness, as a function of how many of the first ten votes were in-network votes, i.e., came from fans of the previous voters. There is a clear inverse relationship between interestingness and the fraction of in-network votes, and this relationship is already visible within the first ten votes. We define a story to be interesting if it receives at least 520 votes,

and not interesting if it received fewer than 520 votes (shown as a dotted line in Figure 4).² As shown in our previous work, many of the stories submitted by top users (who are also well connected) were deemed to be uninteresting by Digg’s audience, receiving comparatively few votes. On the other hand, almost all of the stories submitted by poorly connected users were found to be highly interesting, with many gathering thousands of votes. One of the exceptions was a story that earned only 185 votes. One of the early voters for this story was kevinrose, the founder of Digg and the user with the largest number of fans. The extra visibility that kevinrose’s vote gave to the story, helped promoted this uninteresting story to the front page.

The inverse relationship between interestingness and story’s cascade is especially significant for top users, ones who were responsible for multiple front page submissions. Figure 4 shows this for three select top users. Generally, stories that garnered fewer votes from within the social network of the first ten voters were deemed to be more interesting, as indicated by their higher final vote count.

These observations suggest that there are two mechanisms for the spread of interest in a story on Digg: interest-based and network-based. A highly interesting story will spread from many independent seed sites, as users unconnected to the voters discover it with some small probability and propagate interest in it to their own fans. A story that is interesting to a narrow community, however, will spread within this community only, without being picked up by unconnected users.

4.2 Predicting interestingness

The evidence presented in the section above suggests that it is possible to predict how interesting a story is by monitoring how interest in it spreads through the social network. Moreover, it should be possible to make the prediction relatively early, after the first ten votes. Digg generally waits longer, until a story accumulates at least 40 votes. Such prediction is especially useful for stories submitted by top users who tend to have bigger and more active social networks.

We trained a C4.5 (J48) decision tree classifier [23] on 207 stories in our dataset. Each story had three attributes: number of in-network votes within the first ten votes (v_{10}), number of users watching the submitter ($fans_1$) and a boolean attribute indicating whether the story was interesting or not. The story was judged interesting if it received more than 520 votes. Figure 5 shows the learned decision tree. Results of 10-fold validation indicate that this tree correctly classifies 174 of the examples, and misclassifies 33 examples.

We tested the learned model on stories extracted from the upcoming queue on June 30, 2006. This dataset consisted of 900 stories submitted within the same time period as the

²This threshold is somewhat arbitrary, but was chosen based on Figure 1(a), which indicated that 20% of the stories in the sample received fewer than 500 votes, suggesting 500 as the interestingness threshold. Two stories in our sample that were submitted by top users were close to this threshold, with 505 and 507 votes (with five in-network votes each). We made the decision to raise the interestingness threshold to 520 and keep these ambiguous cases in the sample.

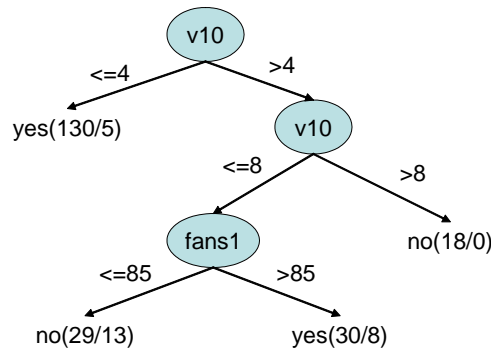


Figure 5: Decision tree classifier trained on the votes data.

data analyzed above, but not yet promoted to the front page. We augmented this data by retrieving the final number of votes received by stories. From this set, we kept only the stories that were submitted by top users (with rank ≤ 100) and received at least 10 votes, leaving 48 stories.

We used the learned classifier in Figure 5 to predict whether a story was interesting (received more than 520 votes). The classifier correctly predicted 36 examples (TP=4, TN=32) and made 12 errors (FP=11, FN=1).³ It is difficult to compare the predictions made by our algorithm to those made by Digg, because some of the stories that Digg did not promote could have ended up receiving many votes and being deemed interesting. When we limit the comparison only to the stories that Digg did promote, of the 14 stories promoted by Digg, only five went on to receive more than 520 votes ($P=TP/(TP+FP)=0.36$), in other words, were judged as interesting by the Digg community. In contrast, our algorithm said that seven of these stories were interesting, and of these four received more than 520 votes ($P=0.57$).

5. RELATED WORK

Recently there has been a lot of interest in utilizing social interactions for word-of-mouth advertising of products. Instead of targeting customers indiscriminately, this type of *viral marketing* [5, 19, 13] aims at choosing certain *influential* nodes in the network that have the potential to influence many others. From the algorithmic standpoint, the problem is to find the optimal set of nodes which will maximize the influence propagation in the network [5, 10]. Clearly, this problem is different from the one considered here, as we are interested in describing the collective behavior rather than affecting it. However, we would like to note that there has been some anecdotal evidence that certain companies have tried to *recruit* influential users to post favorable stories [21].

Other researchers have used Digg’s trove of empirical data to study dynamics of voting. Wu and Huberman [25] showed that the distribution of votes received by a large number of front page stories can be described by a simple stochastic model, parameterized by a single quantity that characterizes the rate of decay of interest in a news article. They

³The notation denotes true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN).

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. The problem studied by Wu and Huberman is complementary to ours, as they studied dynamics of stories *after* they hit the front page, and they do not identify a mechanism for the spread of interest in a story. We, on the other hand, propose, and empirically study, social networks as a mechanism for the spread of interest in a story. We plan to explore the connection between the novelty parameter of Wu and Huberman and story interestingness in our framework. Crane and Sornette [3] analyzed a large number of videos posted on YouTube. By looking at the dynamics of the number of votes received by the videos, they found that they could identify high quality videos, whether they were selected by YouTube editors, or spontaneously became popular. Like Wu and Huberman, they looked at aggregate statistics, not the microscopic dynamics of the spread of interest in stories.

6. CONCLUSION

We studied empirically the spread of interest in news stories on the social news aggregator Digg. We found that social networks play a significant role in promoting stories. In addition, we show that the pattern of social voting can be used for predicting how interesting the story will be. Although our study was carried out on data from Digg, we believe that its conclusions will apply to other social media sites that use social networks to promote content.

As a future work, it will be interesting to analyze more thoroughly the role of network's structural properties on the voting dynamics. Indeed, it is known that structural properties can have a significant impact on various dynamical processes on networks. For instance, it is known that power-law degree distribution observed in many real-world networks can lead to vanishing threshold for epidemics [18, 17] for certain models, in a sharp contrast with the results for random Erdos-Renyi networks. Furthermore, the presence of well-connected clusters of nodes can impact the transient dynamics of various influence propagation models[6]. This latter phenomenon can be especially important in networks with well-defined *community structure* [7, 14].

7. REFERENCES

- [1] M. Arrington. Troubles in diggville. <http://www.techcrunch.com/2006/09/06/troubles-in-diggville/,09/06/2006>.
- [2] M. Calore. Digg fights top users for control. *Wired News*, 09/07/2006.
- [3] R. Crane and D. Sornette. Viral, quality, and junk videos on youtube: Separating content from noise in an information-rich environment. In *Proc. of AAAI symposium on Social Information Processing*, Menlo Park, CA, 2008. AAAI.
- [4] Pedro Domingos and Matt Richardson. Mining the network value of customers. In *KDD '01: Proc. of the 7th ACM SIGKDD Int. Conf. on Knowledge discovery and data mining*, pp. 57–66, New York, NY, USA, 2001.
- [5] Pedro Domingos and Matt Richardson. Mining the network value of customers. In *KDD '01: Proc of the seventh ACM SIGKDD Int Conf on Knowledge discovery and data mining*, pp. 57–66, New York, NY, USA, 2001.
- [6] Aram Galstyan and Paul Cohen. Cascading dynamics in modular networks. *Phys. Rev. E*, 75:036109, 2007.
- [7] M. Girvan and M. E. Newman. Community structure in social and biological networks. *Proc Natl Acad Sci U S A*, 99(12):7821–7826, June 2002.
- [8] Mark S. Granovetter. The strength of weak ties. *The American Journal of Sociology*, 78(6):1360–1380, 1973.
- [9] D. Gruhl, David Liben-Nowell, R. Guha, and A. Tomkins. Information diffusion through blogspace. *SIGKDD Explor. Newsl.*, 6(2):43–52, December 2004.
- [10] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *KDD '03: Proc. of the 9th ACM SIGKDD Int. Conf. on Knowledge discovery and data mining*, pp. 137–146, New York, NY, USA, 2003.
- [11] K. Lerman. Social information processing in social news aggregation. *IEEE Internet Computing*, 11(6):16–28, 2007.
- [12] K. Lerman. User participation in social media: Digg study. In *Proc of WI-AIT workshop Social Media Analysis*, 2007.
- [13] Jure Leskovec, Lada A. Adamic, and Bernardo A. Huberman. The dynamics of viral marketing. In *EC '06: Proc of the 7th ACM Conf on Electronic commerce*, pp. 228–237, New York, NY, USA, 2006.
- [14] M. E. J. Newman. Modularity and community structure in networks. *Proc Natl Acad Sci USA*, 103(23):8577–8582, June 2006.
- [15] M.E.J. Newman. The spread of epidemic disease on networks. *Phys. Rev.*, E 66:016128, 2002.
- [16] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Phys. Rev. Letters*, 86(14):3200–3203, 2001.
- [17] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic dynamics and endemic states in complex networks. *Phys. Rev. E*, 63(6):066117, 2001.
- [18] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.*, 86(14):3200–3203, 2001.
- [19] Matthew Richardson and Pedro Domingos. Mining knowledge-sharing sites for viral marketing. In *KDD '02: Proc of the 8th ACM SIGKDD Int Conf on Knowledge discovery and data mining*, pp. 61–70, New York, NY, USA, 2002.
- [20] K. Rose. Digg friends. <http://diggtheblog.blogspot.com/2006/09/digg-friends.html,09/2006>.
- [21] K. Rose. A couple of updates. <http://blog.digg.com/?p=60,02/01/2007>.
- [22] 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.
- [23] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, 2nd edition, 2005.
- [24] F. Wu, B. Huberman, L. Adamic, and J. Tyler. Information flow in social groups. *Physica A*, 2003.
- [25] F. Wu and B. A. Huberman. Novelty and collective attention. *PNAS* 104:45, 17599–17601, 2007.