

Information Contagion: an Empirical Study of the Spread of News on Digg and Twitter Social Networks

Kristina Lerman and Rumi Ghosh

USC Information Sciences Institute
Marina del Rey, CA 90292, USA

Abstract

Social networks have emerged as a critical factor in information dissemination, search, marketing, expertise and influence discovery, and potentially an important tool for mobilizing people. Social media has made social networks ubiquitous, and also given researchers access to massive quantities of data for empirical analysis. These data sets offer a rich source of evidence for studying dynamics of individual and group behavior, the structure of networks and global patterns of the flow of information on them. However, in most previous studies, the structure of the underlying networks was not directly visible but had to be inferred from the flow of information from one individual to another. As a result, we do not yet understand dynamics of information spread on networks or how the structure of the network affects it. We address this gap by analyzing data from two popular social news sites. Specifically, we extract social networks of active users on Digg and Twitter, and track how interest in news stories spreads among them. We show that social networks play a crucial role in the spread of information on these sites, and that network structure affects dynamics of information flow.

Introduction

Social scientists have long recognized the importance of social networks in the spread of information (Granovetter 1973) and innovation (Rogers 2003). Modern communications technologies, notably email and more recently social media, have only enhanced the role of networks in marketing (Domingos and Richardson 2001; Kempe, Kleinberg, and Éva Tardos 2003), information dissemination (Wu et al. 2004; Gruhl and Liben-nowell 2004), search (Adamic and Adar 2005), and expertise discovery (Davitz et al. 2007). The recent DARPA Network Challenge¹ successfully tested the ability of online social networks to mobilize massive ad-hoc teams to solve real-world problems, which could potentially improve disaster response and coordination of relief efforts. In addition to making social networks ubiquitous, social media sites have given researchers access to massive quantities of data for empirical analysis. These data sets offer a rich source of evidence for studying the structure of social

networks (Leskovec and Horvitz 2008) and the dynamics of individual (Vázquez et al. 2006) and group behavior (Hogg and Lerman 2009), efficacy of viral product recommendation (Leskovec, Adamic, and Huberman 2006), global properties of the spread of email messages (Wu et al. 2004; Liben-Nowell and Kleinberg 2008) and blog posts (Leskovec et al. 2007b), and identification of influential blogs (Gruhl and Liben-nowell 2004; Leskovec et al. 2007a). In most of these studies, however, the structure of the underlying network was not visible but had to be inferred from the flow of information from one individual to another. This posed a serious challenge to our efforts to understand how the structure of the network affects dynamics of information spread on it.

Understanding this question is especially critical for the effective use of social media and peer production systems, which often aggregate over activities of, or contributions made by, many people in order to identify trending topics and noteworthy contributions. Most of these sites also highlight activities of a person's social network links. Since people create links to others who are similar to them, or whose contributions they find interesting, the dynamics of information on a social network may be different from its dynamics within the general population. Separating in-network from out-of-network activity allows us, among other things, to better estimate the inherent quality of the contributions (Crane and Sornette 2008) or predict their future activity (Hogg and Lerman 2010; Lerman and Galstyan 2008). This will in turn allow us to separate high quality contributions from noise.

Social news sites Digg and Twitter offer a unique opportunity to study dynamics of information spread on social networks. Both sites have become important sources of timely information for people. The social news aggregator Digg allows users to *submit* links to news stories and *vote* on stories submitted by other users. On the microblogging service Twitter users *tweet* short text messages that often contain links to news stories and comment on or *retweet* messages of others. Both sites enable users to explicitly create links to other users they want to follow. Another important common feature is data transparency, with both sites providing programmatic access to detailed data about story and user activity.

This paper presents an empirical study of the role of so-

cial networks in the spread of information on Digg and Twitter. For our study we collected data about popular stories on Digg and Twitter that includes information about who voted or retweeted the story and when. In addition, we extracted the social networks of active users on these sites. These data sets allow us to empirically characterize individual dynamics, network structure, and to map the spread of interest in news stories through the network. First, we empirically characterize the structure of social networks on both sites. While the number of fans a user has on each site exhibits a long-tail distribution, Digg's social network is denser and more interconnected than Twitter's, as judged by the number of reciprocated links and the network clustering coefficient. We also show that user activity on both sites has a power-law distribution, albeit with different exponents. Next, we study evolution of the number of votes stories receive. We show that user interface affects dynamics of votes, with evolution of Digg stories going through two distinct stages. Nevertheless, the number of votes accumulated by stories on both sites saturates after a period of about a day to a value that reflects their popularity. Next, we study how information spreads through the social network by measuring how the number of in-network votes a story receives, i.e., votes from fans of the submitter or previous voters, changes in time. We show that the structure of the network affects dynamics of information spread, with information reaching nodes faster in a denser network of Digg than Twitter. However, Twitter stories spread farther, as judged by the total number of in-network votes they receive. We conclude with a discussion of implications of the study.

Social News

Social media has become an important channel for people to share information. On Digg, Twitter, Slashdot, Reddit, and Facebook, among others, users post news or links to news stories, discuss them, and share their opinions in real time. Often, these sites are the first to break important news. After the Christmas 2009 failed attempt to blow up a US commercial airliner, Twitter was the first source to report new security measures for international flights (Carr 2010). In addition to news, these sites are being used as a tool to organize people. For example, in the aftermath of the disputed elections in Iran in June 2009, the opposition movement used Twitter to mobilize the public, organize protests, and inform people about the latest developments, which was more vital in the absence of reliable official information sources.

Digg (<http://digg.com>) is a popular social news aggregator with over 3 million registered users. Digg allows users to submit links to and rate news stories by voting on, or *digg*ing, them. There are many new submissions every minute, over 16,000 a day. Digg picks about a hundred stories daily to feature on its front page. Although the exact promotion mechanism is kept secret, it appears to take into account the number and the rate at which story receives votes. Digg's success is largely fueled by the emergent front page, created by the collective decisions of its many users.

A newly submitted story goes to the *upcoming* stories list, where it remains for 24 hours, or until it is promoted to the *front page*, whichever comes first. Newly submitted stories

are displayed as a chronologically ordered list, with the most recent story at the top of the list, 15 stories to a page. Promoted (or 'popular') stories are also displayed in a reverse chronological order on the front pages, 15 stories to a page, with the most recently promoted story at the top of the list. The importance of being promoted has, among other things, spawned a black market² which claims the ability to manipulate the voting process.

Digg also allows users to designate friends and track their activities. The *friends interface* allows users to see the stories friends recently submitted or voted for. The friendship relationship is asymmetric. When user *A* lists user *B* as a *friend*, *A* can watch the activities of *B* but not vice versa. We call *A* the *fan* of *B*. A newly submitted story is visible in the upcoming stories list, as well as to submitter's fans through the friends interface. With each vote it also becomes visible to voter's fans. The friends interface can be accessed by clicking on *Friends Activity* tab at the top of any Digg page. In addition, a story submitted or voted on by user's friends receives a green ribbon on the story's Digg badge, raising its visibility to fans.

We used Digg API to collect data about 3,553 stories promoted to the front page in June 2009. The data associated with each story contained story title, story id, link, submitter's name, submission time, list of voters and the time of each vote, the time the story was promoted to the front page. In addition, we collected the list of voters' friends. From this information, we were able to reconstruct the fan network of Digg users who were active during the sample period.

Twitter (<http://twitter.com>) is a popular social networking site that allows registered users to post and read short (at most 140 characters) text messages, which may contain URLs to online content, usually shortened by a URL shortening service such as bit.ly or tinyurl. A user can also retweet or comment on another user's post, usually prepending it with a string "RT @*x*," where *x* is a user's name. Posting a link on Twitter is analogous to submitting a new story on Digg, and retweeting the post is analogous to voting for it. Like Digg, Twitter allows users to designate as friends other users whose posts they want to follow. Being a *follower* on Twitter is equivalent to being a fan on Digg.

Twitter restricts large-scale access to its data to a limited number of entities. One of these, Tweetmeme (<http://tweetmeme.com>), aggregates all Twitter posts to determine frequently retweeted URLs, categorizes the stories these URLs point to, and presents them as news stories in a fashion similar to Digg's front page. We collected data from Tweetmeme using specialized page scrapers developed using Fetch Technologies's AgentBuilder tool. For each story, we retrieved the name of the user who posted the link to it, the time it was posted, the number of times the link was retweeted, and details of up to 1000 of the most recent retweets. For each retweet, we extracted the name of the user, the text and time stamp of the retweet. We were limited to 1000 most recent retweets by the structure of Tweetmeme. We extracted 398 stories from Tweetmeme that were originally posted between June 11, 2009 and July 3, 2009. Of

²As an example, see <http://subvertandprofit.com>

these, 329 stories had fewer than 1000 retweets. Next, we used Twitter API to download profile information for each user in the data set. The profile included the complete list of user’s friends and followers.

Characteristics of User Activity

We define as *active user* any user who voted for at least one story on Digg or retweeted at least one story on Twitter. There are 139,409 active Digg and 137,582 active Twitter users in our sample. On Digg, 71,834 active users designated at least one other user as a friend, with a total of 258,220 friend links. Active users on Twitter were connected to 6,200,051 users. From this data, we were able to reconstruct the fan networks of active users, i.e., active users who are watching activities of other users. Figure 1 shows the distribution of number of active fans and followers per user. Digg’s distribution, shown in Fig. 1(a), has a long-tail shape that is common to degree distributions in real-world complex networks (Clauset, Shalizi, and Newman 2009). Twitter’s distribution, shown in Fig. 1(b), has a peak at around 100 followers and a long tail.

As the numbers above suggest, the Digg social network is denser, more tightly knit than the Twitter social network. We measure density by the number of reciprocal friendship links and the modified clustering coefficient. A reciprocal, or mutual, friendship link exists when user A marks B as friend and vice versa. There were 125,219 such links among 279,725 distinct users in the Digg sample and 3,973,892 mutual links among 6,200,051 users in the Twitter sample. Normalizing these counts by the number of all possible mutual links in the network gives us the fraction of mutual links f_m . For Digg $f_m = 3.20 \times 10^{-6}$, and for Twitter $f_m = 2.07 \times 10^{-7}$, an order of magnitude smaller. The clustering coefficient f_c measures the degree to which a node’s network neighbors are interlinked. We define the clustering coefficient for directed networks such as those that exist on Digg and Twitter as the fraction of closed triangles that exist out of all possible sets of three nodes, or triples. For simplicity, we define a closed triangle as a cycle of length three that exists when A lists B as a friend, B lists C and C lists A as a friend. There were 166,239 such triangles in the Digg network, giving us the clustering coefficient $f_c = 7.60 \times 10^{-12}$, and 4,566,952 triangles on Twitter, giving the clustering coefficient of $f_c = 1.92 \times 10^{-14}$ that is two orders of magnitude smaller. Due to the size of the networks, we implemented these metrics using Hadoop³. We suspect that the differences in density of the two networks are due to their age, since Twitter is a more recent service than Digg. With time, we expect the Twitter network to grow denser (Leskovec, Kleinberg, and Faloutsos 2005) and become as tightly knit as Digg.

Next, we characterize users’ voting activity. The 139,409 active users in the Digg data set cast 3,018,197 votes on 3,553 stories. User activity is not uniform, as shown in inset Fig. 1(a). While majority of users cast fewer than 10 votes, some users voted on thousands of stories over the sample

time period. The distribution of the number of retweets per user in the Twitter data set has a similar shape, with the number of retweets per user ranging from 1 to about 100. The difference in slopes in these distribution is likely explained by the level of effort (Wilkinson 2008) required to vote on Digg vs retweet on Twitter.

Dynamics of Voting

Our data sets contain a complete record of voting on Digg front page stories and frequently retweeted stories on Twitter. From this data we can reconstruct dynamics of voting. In addition to voting history, we also know the active fan network of Digg and Twitter users and use this information to check whether a particular voter is a fan of the submitter or previous voters. We call these in-network votes *fan votes*. This information allows us to study how interest in the story spreads through the social networks on Digg and Twitter.

Figure 2(a) shows the evolution of the number of votes received by three Digg stories about post-election unrest in Iran in June 2009. While the details of the dynamics differ, the general features of votes evolution are shared by all Digg stories and can be described by a stochastic model of social voting (Hogg and Lerman 2009). While in the upcoming stories queue, a story accumulates votes at some slow rate. The point where the slope abruptly changes corresponds to promotion to the front page. After promotion the story is visible to a large number of people, and the number of votes grows at a faster rate. As the story ages, accumulation of new votes slows down (Wu and Huberman 2007) and finally saturates. Figure 2(b) shows the evolution of the number of times stories on the same topics were retweeted. The number of retweets grows smoothly until it saturates. It takes about a day for the number of votes/retweets to saturate on both sites.

Distribution of popularity The total number of times the story was voted for and retweeted reflects their popularity among Digg and Twitter users respectively. The distribution of story popularity on either site, Figure 3, shows the ‘inequality of popularity’ (Salganik, Dodds, and Watts 2006), with relatively few stories becoming very popular, accruing thousands of votes, while most are much less popular, receiving fewer than 500 votes.⁴ The most common number of votes by a story is around 500 on Digg and 400 on Twitter. These values are well described by a lognormal distribution (shown as the red line in the figure).

The log-normal distribution of story popularity is typical of the “heavy-tailed” distributions associated with social production and consumption of content. In a heavy-tailed distribution a small but non-vanishing number of items generate uncharacteristically large amount of activity. These distributions have been observed in a variety of contexts, including voting on Digg (Wu and Huberman 2007) and Essembly (Hogg and Szabo 2009), edits of

³<http://hadoop.apache.org/>

⁴This distribution applies to Digg’s front page stories only. Stories that are never promoted to the front page receive very few votes, in many cases just a single vote from the submitter.

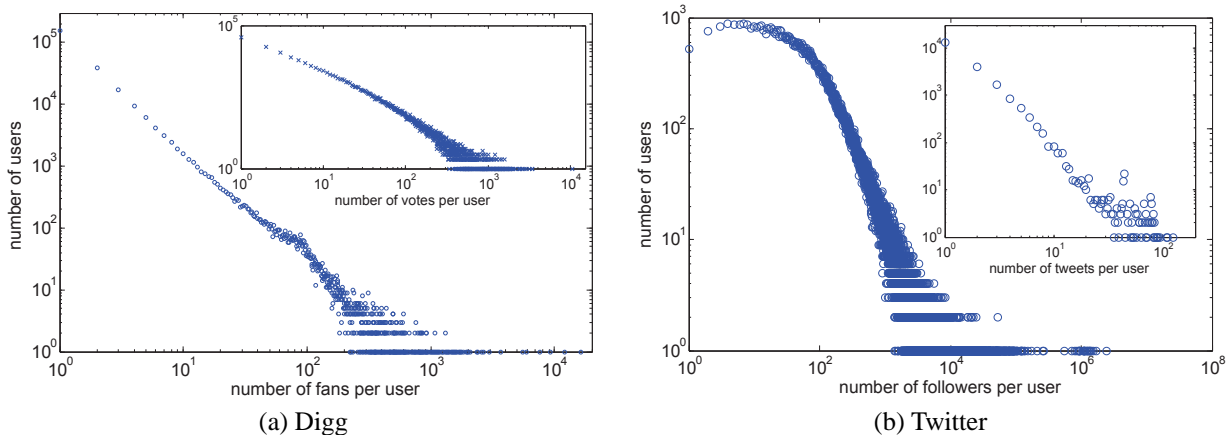


Figure 1: Distribution of user activity. (a) Number of active fans per user in the Digg data set vs the number of users with that many fans. Inset shows distribution of voting activity, i.e., number of votes per user vs number of users who cast that many votes. (b) Number of active followers per user in the Twitter data set vs the number of users with that many followers. Inset shows distribution of retweeting activity.

Wikipedia articles (Wilkinson 2008), and music downloads (Salganik, Dodds, and Watts 2006). Understanding the origin of such distributions is the next challenge in modeling user activity on social media sites.

Dynamics of Voting on Networks

At the time of submission, a Digg story is visible on the upcoming stories list and to submitter’s fans through the friends interface. As users vote on the story, it becomes visible to their own fans via the friends interface. Analogous to the spread of a contagious disease (Newman 2002), interest in the story cascades through the social network. When the story is promoted to the front page, it becomes visible to many nonfans, although users are still able to pick out stories their friends liked through the green ribbon on the story’s Digg badge. Similarly, a new post on Twitter is visible to submitter’s followers, and every user who retweets the story broadcasts it to her own followers. Although aggregators like Tweetmeme attempt to identify popular stories on Twitter in Digg-like fashion, there is no evidence that they boost their visibility to nonfans.

We can trace the cascade of interest in a story through the underlying social network of Digg (Twitter) by checking whether a new vote (retweet) came from a fan (follower) of any of the previous voters, including the submitter. We call such votes or retweets *fan votes*, regardless of whether we are talking about Digg or Twitter. Therefore, the cascade (“information contagion” in the title of this article) starts with story’s submitter and grows as the story accrues fan votes. Researchers have studied information cascades in email chain letters (Wu et al. 2004; Liben-Nowell and Kleinberg 2008) and blog posts (Gruhl and Liben-nowell 2004; Leskovec et al. 2007b) in order to obtain insights into the structure of the network, identify influential nodes within it, or predict popularity of content (Lerman and Galstyan 2008). Characterizing in-

formation cascades is necessary for creating a model of the dynamics of information on networks.

Dynamics and distribution of fan votes The dashed lines in Figure 2 show how the number of fan votes received by each story, grows in time. Their evolution is similar to that of all votes, and growth saturates after a period of about a day. The value at which growth saturates shows the story’s range, or how widely it penetrates the social network. Figure 4 shows the distribution of cascade sizes generated by Digg and Twitter stories. These distributions are markedly different from the distribution of story popularity shown in Fig. 3. Although the distribution of network cascades of Digg stories, Fig. 4(a), is slightly asymmetrical, it is best described by a normal with the mean and standard deviation equal to 104.27 and 32.31 votes respectively, not the log-normal distribution in Fig. 3(a). It is also unlike distribution of cascade sizes in a blog post network, which has a power law distribution (Leskovec et al. 2007b). Remarkably, there are no stories that did not generate a cascade, i.e., which did not receive any fan votes.

The inset in Figure 4(a) shows the distribution of votes from submitter’s fans only. It is also described by a normal function with a mean around 50 votes. A small fraction of stories, fewer than 400, did not have any votes from submitter’s fans. This indicates that active users who are fans of the submitter are also fans of other voters, i.e., that the social network of active Digg users is dense and highly interlinked. This observation is supported by the finding of a relatively high clustering coefficient of the Digg social network.

The distribution of cascade sizes of Twitter stories is shown in Fig. 4(b). These also appear to be normally distributed, although a substantial number of stories do not spread on the network. This distribution is broader than that of Digg stories, which indicates that stories spread farther on the Twitter network. The distribution of the number of votes cast by submitter’s followers, shown in inset in Fig. 4(b),

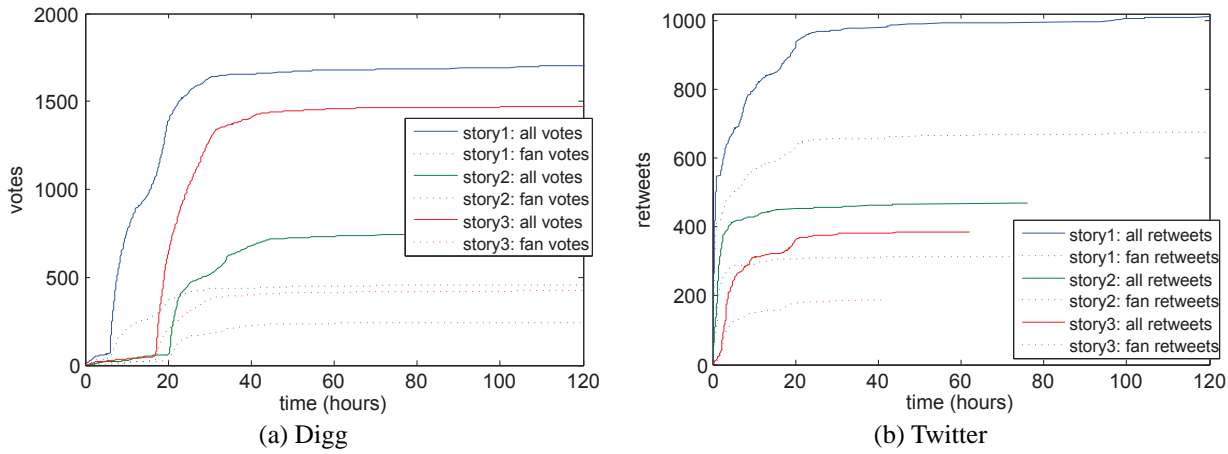


Figure 2: Dynamics of stories on Digg and Twitter. (a) Total number of votes (diggs) and fan votes received by stories on Digg since submission. (b) Total number of times a story was retweeted and the number of retweets from followers since the first post vs time. The titles of stories on Digg were: story1: “U.S. Government Asks Twitter to Stay Up for #IranElection”, story2: “Western Corporations Helped Censor Iranian Internet”, story3: “Iranian clerics defy ayatollah, join protests.” The titles of retweeted stories were: story1: “US gov asks twitter to stay up”, story2: “Iran Has Built a Censorship Monster with help of west tech”, story3: “Clerics join Iran’s anti-government protests - CNN.com.”

is markedly different from Digg. The vast majority of the stories did not receive any votes from submitter’s followers, indicating that submitter’s and other voters’ followers are disjoint. This observation is supported by our finding that the Twitter social network is sparsely interconnected.

Evolution of fan votes Figure 5 shows how the number of fan votes (size of the cascade), aggregated over all stories, grows during the early stages of voting or retweeting. While there is significant variation in the number of fan votes received by a story, the aggregate exhibits a well-defined trend. The solid black lines show the median cascade size, while thin gray lines show the envelope of the boundary that is one standard deviation from the mean.

The cascade grows steadily with new votes on Digg (Fig. 5(a)), although faster initially, indicating that there are two distinct mechanisms for story visibility on Digg. This is seen more clearly in Fig. 5(b), which shows the probability that next vote is a fan vote and will increase the size of the cascade. We separate votes cast before promotion from those cast after the story is promoted. Before promotion, this probability is almost constant, at $p = 0.74$. After promotion, it decays to a lower, but also almost constant value $p = 0.3$. This is consistent with our hypothesis that before promotion social networks are the primary mechanism for spreading interest in new stories. Although a story is also visible on the upcoming stories list, few users actually discover stories there. With 16,000 daily submissions, a new story is quickly submerged by new submissions and is pushed to page 15 of the upcoming stories list within the first 20 minutes. Few users are likely to navigate that far (Huberman et al. 1998). Promotion to the front page, which generally happens when a story accrues between 50 and 100 votes, exposes the story to a large and diverse audience, making social networks less of a factor in its spread, since large numbers of Digg users

who read front page stories do not befriend others.

The spread of interest in stories through the Twitter network, shown in Figure 5(c), is similar to Digg. As on Digg, the median number of fan votes rises steadily during the early stages of voting. However, the rate of growth is nearly constant, indicating there is a single significant mechanism for making stories visible to voters, namely the social network. The probability that next retweet is from a fan, shown in Fig. 5(d), rises slowly from around $p = 0.4$ to $p = 0.55$. This value is lower than pre-promotion probability of next fan vote on Digg. The rate of interest spread appears to depend on the density of network. Initially, Digg stories spread faster through the social network than stories on Twitter, because of Digg’s denser network structure, but after promotion they spread much slower as unconnected users see and vote on the stories.

The dashed lines in Fig. 5(a) & (c) show how the median number of votes from submitter’s fans or followers changes with voting. By the time a story accumulates 50 votes on Digg (at which point some of the stories are promoted to the front page), about half of the votes are from submitter’s fans, and another 10 are from fans of prior voters but not the submitter. After a story receives about 100 votes (by which point most of the stories are promoted), the number of votes from submitter’s fans changes very slowly, while the number of fan votes continues to grow. This indicates that submitter’s fans vote for the story during its early stages and that users pay attention to the stories their friends submit. On Twitter, initial votes are from submitter’s fans, but slows significantly later.

Related Work

Several researchers studied dynamics of information flow on networks, however, empirical studies have produced conflicting results. (Wu et al. 2004) examined patterns of email

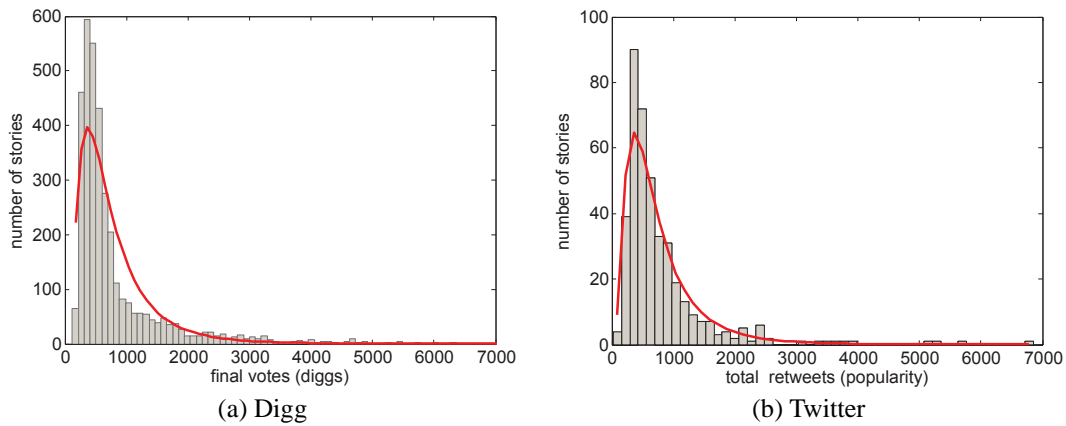


Figure 3: Distribution of story popularity. (a) Distribution of the total number of votes received by Digg stories, with line showing log-normal fit. The plot excludes the 15 stories that received more than 6,000 votes. (b) Distribution of the total number of times stories in the Twitter data set were retweeted, with the line showing log-normal fit.

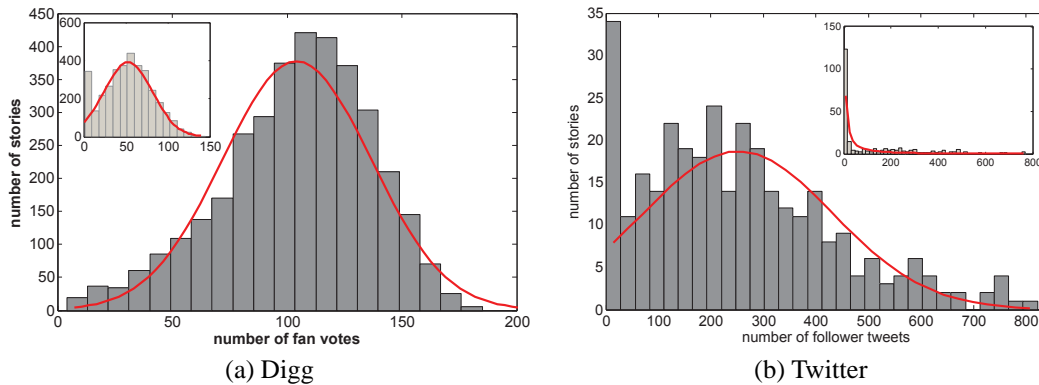


Figure 4: Distribution of story cascade sizes. (a) Histogram of the distribution of the total number of fan votes received by Digg stories (size of the interest cascade). The inset shows the distribution of the number of votes from submitter’s fans. (b) Histogram of the distribution of the total number of retweets from followers. The inset shows the distribution of the number of retweets of a story from submitter’s followers.

forwarding within an organization and found that email forwarding chains terminate after an unexpectedly small number of steps. They argued that unlike the spread of a virus on a social network, which is expected to reach many individuals, the flow of information is slowed by decay of similarity among individuals within the social network. They measured similarity by distance in organizational hierarchy between the two individuals within an organization, or in general, as a number of edges separating two nodes within a graph. Similarly, in a large-scale study of the effectiveness of word-of-mouth product recommendations, (Leskovec, Adamic, and Huberman 2006) found that most recommendation chains terminate after one or two steps. However, authors noted sensitivity of recommendation to price and category of product, leaving open the question whether social networks are an effective tool for disseminating information, rather than purchasing products. Contrary to these studies, we find that information, such as news, reaches many individuals within a social network.

Moreover, the reach of information spread does not seem to depend on similarity between users, at least when similarity is measured by number of edges between them. On Digg, whose users are highly interconnected, a story does not reach as many fans as on Twitter, where users are less densely connected.

Like Wu et al., (Liben-Nowell and Kleinberg 2008) studied the patterns of forwarding of two popular email petitions. Unlike their expectations, the forwarding chains produced long narrow, rather than bushy wide, trees. In these studies, however, the structure of the underlying social network was not directly visible but had to be inferred by observing new signatures on the forwarded petitions. This method offers only a partial view of the network and does not identify all edges between individuals that participated in the email chain. If an individual has already forwarded the message, she will not do so again, and an edge between her and the sender will not be observed. In our study, on the other hand, the networks are extracted independently of data about the

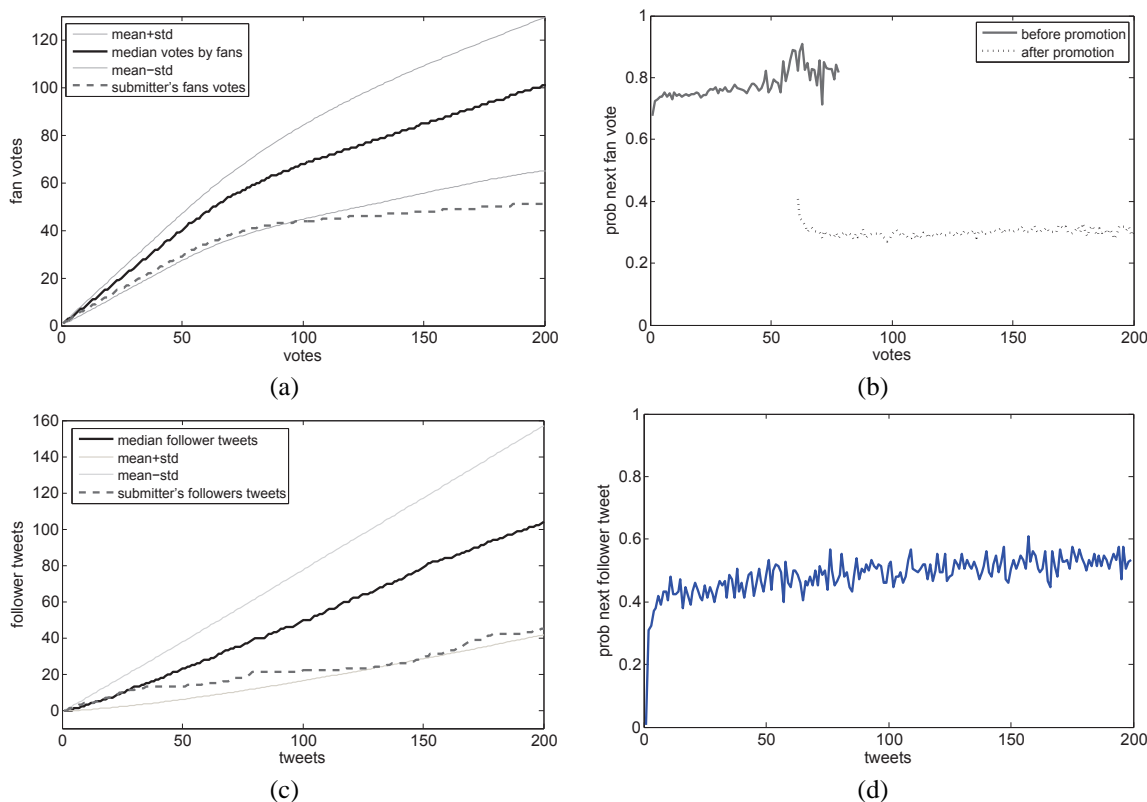


Figure 5: Spread of interest in stories through the network. (a) Median number of fan votes vs votes, aggregated over all Digg stories in our data set. Dotted lines show the boundary one standard deviation from the mean. Dashed lines shows the number of votes from fans of submitter. (b) Probability next vote is from a fan before and after the Digg story is promoted. (c) Median number of retweets from followers vs all retweets, aggregated over all stories in the Twitter data set. (d) Probability next retweet is from a follower.

spread of information.

A number of researchers have studied the flow of information and influence in the blogosphere and in a virtual world. (Gruhl and Liben-nowell 2004) traced topic propagation through blogs and used a model of the spread of epidemics on networks (Newman 2002) to characterize the spread of topics through the blogosphere. (Leskovec et al. 2007b) defined an information cascade as a graph of hyperlinks between blog posts. A cascade starts with a cascade initiator, with other blog posts joining the cascade by linking to the initiator or other members of the cascade. Leskovec et al. found that the distribution of cascade sizes follows a power law. In these studies, the networks were derived from the observed links between blog posts, i.e., from the diffusion of information. In our study, on the contrary, they were extracted from the sites independently of data about the diffusion of information. (Bakshy, Karrer, and Adamic 2009) traced the spread of influence in a multi-player online game and found that similar to our findings with social news, influence spreads easily on social networks in virtual worlds. This provides an independent confirmation of the importance of social networks in the dynamics of information flow.

Conclusion

We conducted an empirical analysis of user activity on Digg and Twitter. Though the two sites are vastly different in their functionality and user interface, they are used in strikingly similar ways to spread information. First, on both sites users actively create social networks by designating as friends others whose activities they want to follow. Second, users employ these networks to discover and spread information, including news stories. The mechanism for the spread of information is the same on both sites, namely, users watch their friends' activities — what they tweet or vote for — and by their own tweeting and voting actions they make this information visible to their own fans or followers. In spite of the similarities, there are quantitative differences in the structure and function of social networks on Digg and Twitter. Digg networks are dense and highly interconnected. A story posted on Digg initially spreads quickly through the network, with users who are following the submitter also likely to follow other voters. After the story is promoted to Digg's front page, however, it is exposed to a large number of unconnected users. The spread of the story on the network slows significantly, though the story may still generate a large response from Digg audience. The Twitter social network is less dense than Digg's, and stories spread through

the network slower than Digg stories do initially, but they continue spreading at this rate as the story ages and generally penetrate the network farther than Digg stories.

Understanding characteristics of user activity and the effect social networks have on it will help us make better use of social media and peer production systems. Currently these systems blindly aggregate activities of all users in order to identify high quality contributions. However, since popularity and quality are rarely linked (Salganik, Dodds, and Watts 2006), this method is likely to highlight popular, though trivial, contributions. Separating in-network and out-of-network user activity, however, will lead to a better understanding of social dynamics of peer production systems (Hogg and Lerman 2009; Hogg and Szabo 2009; Lerman and Hogg 2010), which will allow us to better separate high quality contributions from noise (Hogg and Lerman 2010; Crane and Sornette 2008; Lerman and Galstyan 2008).

Acknowledgments

We are grateful to Tad Hogg for valuable insights into data analysis and to Prashant Khanduri for initial analysis of Digg data. This material is based upon work supported by the National Science Foundation under Grant No. 0915678.

References

- [Adamic and Adar 2005] Adamic, L. A., and Adar, E. 2005. How to search a social network. *Social Networks* 27(3):187–203.
- [Bakshy, Karrer, and Adamic 2009] Bakshy, E.; Karrer, B.; and Adamic, L. A. 2009. Social influence and the diffusion of user-created content. In *EC '09: Proc. 10th ACM conference on Electronic commerce*, 325–334.
- [Carr 2010] Carr, D. 2010. Why twitter will endure. *New York Times*.
- [Clauset, Shalizi, and Newman 2009] Clauset, A.; Shalizi, C. R.; and Newman, M. E. J. 2009. Power-law distributions in empirical data. *SIAM Review* 51(4):661+.
- [Crane and Sornette 2008] Crane, R., and Sornette, D. 2008. Viral, quality, and junk videos on youtube: Separating content from noise in an information-rich environment. In *Proc. AAAI symposium on Social Information Processing*.
- [Davitz et al. 2007] Davitz, J.; Yu, J.; Basu, S.; Gutelius, D.; and Harris, A. 2007. ilink: Search and routing in social networks. In *Proc. Knowledge Discovery and Data Mining Conference (KDD-2007)*.
- [Domingos and Richardson 2001] Domingos, P., and Richardson, M. 2001. Mining the network value of customers. In *Proc. KDD*.
- [Granovetter 1973] Granovetter, M. 1973. The strength of weak ties. *The American Journal of Sociology*.
- [Gruhl and Liben-nowell 2004] Gruhl, D., and Liben-nowell, D. 2004. Information diffusion through blogspace. In *Proc. Int. World Wide Web Conference (WWW)*, 491–501.
- [Hogg and Lerman 2009] Hogg, T., and Lerman, K. 2009. Stochastic models of user-contributory web sites. In *Proc. Int. Conference on Weblogs and Social Media*.
- [Hogg and Lerman 2010] Hogg, T., and Lerman, K. 2010. Social dynamics of digg. In *Proc. Int. Conference on Weblogs and Social Media (ICWSM10)*.
- [Hogg and Szabo 2009] Hogg, T., and Szabo, G. 2009. Diversity of user activity and content quality in online communities. In *Proc. Int. Conference on Weblogs and Social Media (ICWSM)*.
- [Huberman et al. 1998] Huberman, B. A.; Pirolli, P. L. T.; Pitkow, J. E.; and Lukose, R. M. 1998. Strong regularities in world Wide Web surfing. *Science* 280(5360):95–97.
- [Kempe, Kleinberg, and Éva Tardos 2003] Kempe, D.; Kleinberg, J.; and Éva Tardos. 2003. Maximizing the spread of influence through a social network. In *KDD '03: Proc. 9th Int. Conf. on Knowledge discovery and data mining*, 137–146.
- [Lerman and Galstyan 2008] Lerman, K., and Galstyan, A. 2008. Analysis of social voting patterns on digg. In *Proc. 1st ACM SIGCOMM Workshop on Online Social Networks*.
- [Lerman and Hogg 2010] Lerman, K., and Hogg, T. 2010. Using a model of social dynamics to predict popularity of online content. In *Proc. 19th Int. World Wide Web Conference*.
- [Leskovec, Adamic, and Huberman 2006] Leskovec, J.; Adamic, L.; and Huberman, B. 2006. The dynamics of viral marketing. In *EC '06: Proc. 7th Conf. on Electronic commerce*, 228–237.
- [Leskovec and Horvitz 2008] Leskovec, J., and Horvitz, E. 2008. Planetary-scale views on a large instant-messaging network. In *WWW '08: Proc. 17th Int. World Wide Web Conference*, 915–924.
- [Leskovec et al. 2007a] Leskovec, J.; Krause, A.; Guestrin, C.; Faloutsos, C.; Vanbriesen, J.; and Glance, N. 2007a. Cost-effective outbreak detection in networks. In *KDD '07: Proc. 13th Int. Conf. on Knowledge discovery and data mining*, 420–429.
- [Leskovec et al. 2007b] Leskovec, J.; McGlohon, M.; Faloutsos, C.; Glance, N.; and Hurst, M. 2007b. Cascading behavior in large blog graphs. In *Proc. 7th SIAM Int. Conference on Data Mining (SDM)*.
- [Leskovec, Kleinberg, and Faloutsos 2005] Leskovec, J.; Kleinberg, J.; and Faloutsos, C. 2005. Graphs over time: densification laws, shrinking diameters and possible explanations. In *KDD '05: Proc. 11th Int. Conf. on Knowledge discovery in data mining*, 177–187.
- [Liben-Nowell and Kleinberg 2008] Liben-Nowell, D., and Kleinberg, J. 2008. Tracing information flow on a global scale using internet chain-letter data. *PNAS* 105(12):4633–4638.
- [Newman 2002] Newman, M. E. J. 2002. Spread of epidemic disease on networks. *Physical Review E* 66(1):016128+.
- [Rogers 2003] Rogers, E. M. 2003. *Diffusion of Innovations, 5th Edition*. Free Press, 5 edition.

- [Salganik, Dodds, and Watts 2006] Salganik, M.; Dodds, P.; and Watts, D. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *Science* 311:854.
- [Vázquez et al. 2006] Vázquez, A.; Oliveira, J. G.; Dezsö, Z.; Goh, K.; Kondor, I.; and Barabási, A. 2006. Modeling bursts and heavy tails in human dynamics. *Phys. Rev. E* 73(3):036127+.
- [Wilkinson 2008] Wilkinson, D. M. 2008. Strong regularities in online peer production. In *EC '08: Proc. 9th Conf. on Electronic commerce*, 302–309.
- [Wu and Huberman 2007] Wu, F., and Huberman, B. A. 2007. Novelty and collective attention. *PNAS* 104(45):17599–17601.
- [Wu et al. 2004] Wu, F.; Huberman, B.; Adamic, L.; and Tyler, J. 2004. Information flow in social groups. *Physica A*.

Information Contagion: an Empirical Study of the Spread of News on Digg and Twitter Social Networks

Kristina Lerman and Rumi Ghosh

USC Information Sciences Institute
Marina del Rey, CA 90292, USA

Abstract

Social networks have emerged as a critical factor in information dissemination, search, marketing, expertise and influence discovery, and potentially an important tool for mobilizing people. Social media has made social networks ubiquitous, and also given researchers access to massive quantities of data for empirical analysis. These data sets offer a rich source of evidence for studying dynamics of individual and group behavior, the structure of networks and global patterns of the flow of information on them. However, in most previous studies, the structure of the underlying networks was not directly visible but had to be inferred from the flow of information from one individual to another. As a result, we do not yet understand dynamics of information spread on networks or how the structure of the network affects it. We address this gap by analyzing data from two popular social news sites. Specifically, we extract social networks of active users on Digg and Twitter, and track how interest in news stories spreads among them. We show that social networks play a crucial role in the spread of information on these sites, and that network structure affects dynamics of information flow.

Introduction

Social scientists have long recognized the importance of social networks in the spread of information (?) and innovation (?). Modern communications technologies, notably email and more recently social media, have only enhanced the role of networks in marketing (?; ?), information dissemination (?; ?), search (?), and expertise discovery (?). The recent DARPA Network Challenge¹ successfully tested the ability of online social networks to mobilize massive ad-hoc teams to solve real-world problems, which could potentially improve disaster response and coordination of relief efforts. In addition to making social networks ubiquitous, social media sites have given researchers access to massive quantities of data for empirical analysis. These data sets offer a rich source of evidence for studying the structure of social networks (?) and the dynamics of individual (?) and group behavior (?), efficacy of viral product recommendation (?), global properties of the spread of email messages (?; ?) and blog posts (?), and identification of influential blogs (?;

?). In most of these studies, however, the structure of the underlying network was not visible but had to be inferred from the flow of information from one individual to another. This posed a serious challenge to our efforts to understand how the structure of the network affects dynamics of information spread on it.

Understanding this question is especially critical for the effective use of social media and peer production systems, which often aggregate over activities of, or contributions made by, many people in order to identify trending topics and noteworthy contributions. Most of these sites also highlight activities of a person's social network links. Since people create links to others who are similar to them, or whose contributions they find interesting, the dynamics of information on a social network may be different from its dynamics within the general population. Separating in-network from out-of-network activity allows us, among other things, to better estimate the inherent quality of the contributions (?) or predict their future activity (?; ?). This will in turn allow us to separate high quality contributions from noise.

Social news sites Digg and Twitter offer a unique opportunity to study dynamics of information spread on social networks. Both sites have become important sources of timely information for people. The social news aggregator Digg allows users to *submit* links to news stories and *vote* on stories submitted by other users. On the microblogging service Twitter users *tweet* short text messages that often contain links to news stories and comment on or *retweet* messages of others. Both sites enable users to explicitly create links to other users they want to follow. Another important common feature is data transparency, with both sites providing programmatic access to detailed data about story and user activity.

This paper presents an empirical study of the role of social networks in the spread of information on Digg and Twitter. For our study we collected data about popular stories on Digg and Twitter that includes information about who voted or retweeted the story and when. In addition, we extracted the social networks of active users on these sites. These data sets allow us to empirically characterize individual dynamics, network structure, and to map the spread of interest in news stories through the network. First, we empirically characterize the structure of social networks on both sites.