Social Informatics: Using Big Data to Understand Social Behavior

Kristina Lerman

Introduction

Modern communications technologies, notably email and more recently social media, have enabled people to interact on an unprecedented scale. The social networks that emerge from these interactions can amplify information (Wu et al. 2004; Gruhl and Liben-nowell 2004), mobilize massive ad-hoc teams (Pickard et al. 2011) and political movements (Lotan et al. 2011), help people discover information (Adamic and Adar 2005; Lerman 2007) and make new connections. In addition to making social networks ubiquitous, social media has given researchers access to massive quantities of data for analysis. These data sets offer a rich source of evidence for studying the structure of networks and the dynamics of individual and group behavior, and ask new questions about social communication. How far and how fast does information spread? How do people respond to new information? What are the mechanisms of information spread and how do individual's cognitive limitations affect them?

We have addressed these questions through a large scale analysis of data from two social media sites: Digg and Twitter. Despite having different functionality and user interface, both sites are used in remarkably similar ways by people to share information with others, thus enabling us to uncover principles of social behavior that generalize across platforms. The social news aggregator Digg allows users to *submit* links to news stories and *recommend* stories submitted by other users by voting for them. On Twitter, users *tweet* short text messages, that often contain links to news stories, or *retweet* messages of others. Both sites allow users to link to others whose activity (i.e., votes and tweets) they want to follow. Upon visiting Twitter, a user is presented with a list of messages most recently tweeted or retweeted by the

K. Lerman (✉)
USC Information Sciences Institute, Marina del Rey, CA, USA
e-mail: lerman@isi.edu

P. Michelucci (ed.), *Handbook of Human Computation*,
DOI 10.1007/978-1-4614-8806-4_60,

followees of the user, i.e., other users whom the given user follows. Similarly, on Digg a user sees a list of news stories recently recommended by those a user follows. By recommending a story, or retweeting a message, in turn, the user acts to further spread the information contained in that story or message.

We trace the flow of information from users to their followers on these sites (using URLs as unique markers of information) and measure its properties. We find that information does not spread to as many people as predicted by a simple model that is commonly used to describe the spread of information. Our attempts to resolve this puzzle illuminates the critical role that individual's limited attention plays in social media.

Social Information Sharing

We studied social information sharing on Digg and Twitter, two popular social media sites for sharing news and other content. For our study, we tracked how items, uniquely identified by URLs, were shared by users. Details of data collection from both sites are described in Lerman et al. (2012).

Figure 1 shows the statistics of social behavior on Digg and Twitter, including the distribution of the number of followers ((a) and (d)) and activity ((b) and (e)), i.e., number of votes or retweets made by each user. While the overwhelming majority of users on both sites shared fewer than ten items (URLs) with followers, a handful of users shared thousands of items over the period of a month. Such heavy-tailed distributions are typical of social production and consumption of content, where a small but non-vanishing number of items generate uncharacteristically large amount of activity, and have been observed in voting on Essembly (Hogg and Szabo 2009), edits of Wikipedia articles (Wilkinson 2008), and music downloads (Salganik et al. 2006) and other and real-world complex networks (Clauset et al. 2009).

The total number of times the URL was shared reflects its popularity. The distribution of popularity on both sites is long-tailed (Fig. 1c, f). It appears that information in social media rarely goes "viral" (Ver Steeg et al. 2011; Goel et al. 2012). The vast majority of items fail to spread at all, reaching only a handful of users. Even the most popular items spread to at most a few thousands users, which is a tiny fraction of the follower graph. Moreover, the distribution of popularity on the two sites is strikingly different: while the distribution of popularity on Digg is well described by a log-normal (shown as the red line), with the mean of 614 votes, there is no preferred popularity for retweeted URLs on Twitter. What gives rise to the difference in distributions of popularity? Wu and Huberman (2007) proposed a phenomenological model that explained the log-normal distribution of popularity on Digg as a byproduct of competition for attention for news stories and their decaying novelty. In contrast, we find that the difference can be explained by Digg's promotion mechanism, which highlights a handful of stories on its popular front page. The test this hypothesis, we gathered statistics about more than 20 K stories submitted to Digg over the course of 1 day in July 2010. The distribution of popularity of these stories is similar to Twitter (Fig. 1c). Of these stories, about 100 were promoted to the front

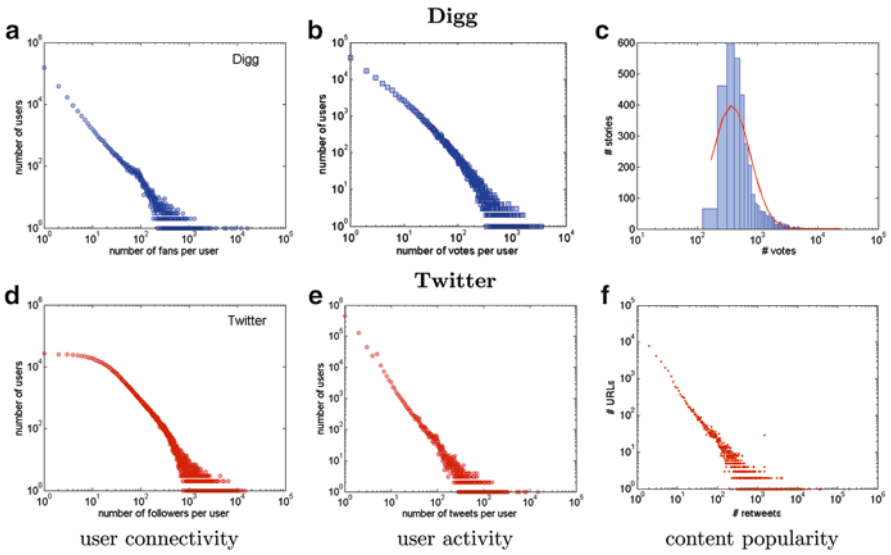

Fig. 1 Characteristics of user activity on Digg and Twitter. Distribution of the number of followers per user on the two sites, distribution of activity, which is given by the number of votes (on Digg) and retweets (on Twitter), and the distribution of popularity of content, as measured by the total votes received by news stories on Digg and the total number of times the URL was retweeted on Twitter. *Red line* in the distribution of votes received by Digg stories is log-normal fit to data

page and their popularity continued to grow. The final popularity of the promoted stories had a log-normal distribution. Therefore, we conclude that the log-normal popularity distribution is a by-product of selection by the promotion algorithm.

A Simple Model of Information Diffusion

Why does some content become popular but not other? How does information spread between people? In order to answer these questions, we need a model of social contagion that describes the microscopic dynamics of the spread of information. One of the simplest such models is the independent cascade model (ICM) (Newman 2022; Kempe et al. 2003; Gruhl and Liben-nowell 2004; Anagnostopoulos et al. 2008), which has been used to describe the spread of a disease in a population (Hethcote 2000). In this model, each exposure of a healthy person by an infected friend leads to an independent chance of the healthy person contracting the disease, and spreading it to her own followers thereby creating a cascade of infections. The likelihood that an exposure leads to an infection is set by pathogen's transmissibility, i.e., how contagious it is. When ICM is stated in the language of information spread, each exposure of a naive individual by an informed friend (e.g., via a tweet), creates an independent chance of information transmission. Therefore, the likelihood that the naive individual becomes informed should increase monotonically with the number of exposures.

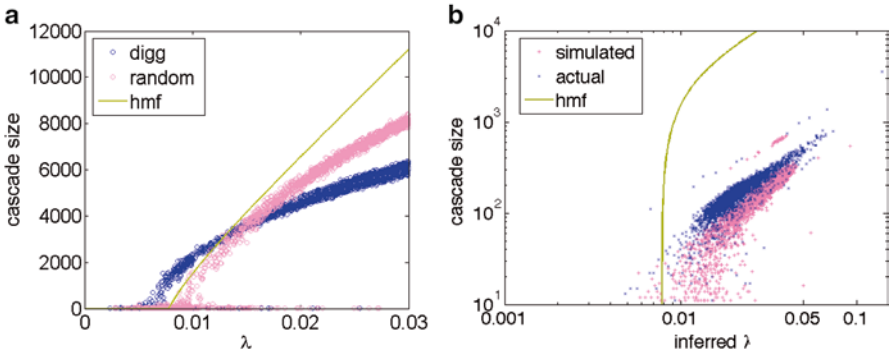

Fig. 2 Cascade size as a function of transmissibility λ. (**a**) Comparison of simulated cascades on the Digg follower graph and on the randomized graph with the same degree distribution. (**b**) Comparison of real and simulated cascades on the Digg graph that are produced using empirical exposure function. Theoretical predictions for a graph of the same size are shown by the *bold line*

Simulations of Information Diffusion

The dynamics of the independent cascade model has been well-studied. Specifically, it is known that there exists a critical value of transmissibility below which the disease does not spread, but above which it reaches a substantial fraction of the population, resulting in an epidemic (Castellano et al. 2009; Satorras and Vespignani 2001; Wang et al. 2003). Moreover, the expected size of an epidemic outbreak of a pathogen with a given transmissibility can be theoretically calculated (Moreno et al. 2002).

Our simulations of the independent cascade model on the Digg follower network confirm these expectations. Starting with random seed node, we generate a cascade as follows (see Ver Steeg et al. (2011) for details). Each time a node is infected, it will attempt to infect each follower independently with probability given by the transmissibility λ. The cascade stops when no new nodes are infected. The number of infected nodes, i.e., cascade size, is shown in Fig. 2, where each point represents a single simulated cascade with transmissibility λ. Dark gray dots represent cascades on the original Digg follower graph, while light gray dots represent cascades on a randomized version of the Digg graph with the same degree distribution. Both curves manifest a critical value of transmissibility, called the *epidemic threshold*, above which cascades spread to a significant fraction of the graph.[1] The location of the epidemic threshold is accurately predicted by the inverse of the largest eigenvalue of the adjacency matrix of the graph (Wang et al. 2003): $\lambda_c^{digg} = 0.00587$ for

[1] Note that even above epidemic threshold, cascades that start in an isolated region of the graph will die out.

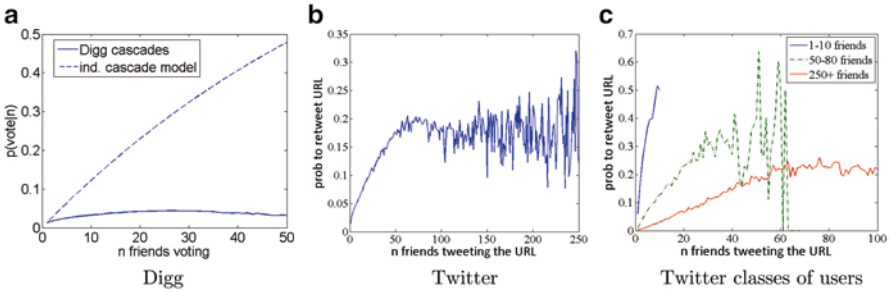

Fig. 3 Response to multiple exposures. (**a**) and (**b**) Show probability of infection given *n* infected friends aggregated over all users on (**a**) Digg and (**b**) Twitter. Plot (**c**) shows exposure response of Twitter users after they are separated into different classes based on their cognitive load, i.e., the number of friends they follow. *Dashed line* in (**a**) show exposure response predicted by the independent cascade model. (**a**) Digg. (**b**) Twitter. (**c**) Twitter classes of users

the original Digg graph and $\lambda_c^{rand} = 0.00928$ for the randomized graph. The size of theoretically predicted cascades is depicted by the gold line, which accurately characterizes both the threshold and growth of cascades on the randomized graph.

Figure 2 presents a puzzle. Information in social media spreads to a far smaller fraction of the population than predicted by the epidemic model. This is not because these URLs have low transmissibility: the Digg dataset, consists of URLs that have been selected for the front page. Nor does it appear to be due to network structure: while structure of the real Digg graph reduces the size of cascades in simulations compared to the randomized graph, it does not suppress it nearly enough to account for the observed sizes of actual outbreaks.

Exposure Response

A potential explanation for why information spread in social media fails to reach epidemic proportions can be found in how people respond to repeated exposures to information, that is, the probability they will rebroadcast the information via a retweet or a vote after multiple friends have tweeted about it or voted for it. According to the independent cascade model, the probability a node becomes infected, e.g., by voting for a story on Digg, increases monotonically with the number of infected neighbors *n* it has. This probability is given by the exposure function:

$$p_{ICM}(\text{ infected} \mid n \text{ infected friends}) = 1 - (1 - \lambda)^n.$$

To measure the exposure function on Digg and Twitter, we isolated users who had exactly *n* infected friends but did not become infected themselves, from users who had *n* infected friends before they themselves became infected. The solid lines in Fig. 3a, b show the probability of Digg and Twitter users respectively to become infected when exposed to information by *n* friends, averaged over all users. Exposure

response on both sites is qualitatively similar. As the number of exposing friends increases, a user's probability to become infected goes up initially, but after a point additional exposure does not further increase response, and may in fact inhibit it. This behavior is similar to adoption of hashtags reported by Romero et al. (2011). In contrast, the dashed line in Fig. 3a depicts exposure response for the independent cascade model. ICM dramatically overestimates infection probability.

When we simulated information diffusion on the Digg follower graph using the empirical exposure response function measured from the data (Ver Steeg et al. 2011), the resulting cascades were dramatically smaller, as shown in Fig. 2b. In fact, the size of simulated cascades (pink dots) is similar to those of real information cascades on Digg (blue dots). It appears that failure to respond to exposures to information stops social epidemics.

Limited Attention in Information Diffusion

We still have a puzzle: why do users fail to respond to repeated exposures by friends? One potential explanation could be that users become "innoculated" to information. In other words, if a user did not find information interesting upon first exposure, she will not find it worthy of spreading upon subsequent exposures. The real explanation is both simpler and more interesting: in a nutshell, users do not see the exposures, and hence do not respond to them.

Our study of how Twitter users respond to messages from friends demonstrated that users are far more likely to retweet a recent message than an old one, and that the more friends a user follows, the less likely he or she is to retweet an older message (Hodas and Lerman 2012). We invoke the concept of limited attention (Kahneman 1973) to explain why people are less likely to retweet older messages. In order to retweet some information, a user first has to find it by wading through a stream of other messages. Reading tweets, however, requires mental effort, of which people have a limited reserve. Limited attention constrains how deeply into his or her stream the user will browse before getting tired, bored or distracted. Since both Twitter and Digg display messages in reverse chronological order, with the most recent message at the top of the screen, the user is far more likely to see recent messages than older ones that are buried deep in their stream. In addition, the more friends the user follows, the faster a message gets buried, and the less likely the user is to see it.

Limited attention alters how well-connected users, i.e., those who follow many others, respond to information. The exposure response functions shown in Fig. 3a, b have been aggregated over all users. These users form a highly heterogeneous group with a wide range of capabilities and motivations to consume and share information. By conflating together behaviors of different types of individuals, heterogeneity may in fact obscure simpler individual behavior (Vaupel and Yashin 1985). Indeed, when we separate users into more homogeneous subpopulations, a different picture emerges. Figure 3c shows the exposure response function of Twitter users who were

separated into subpopulations based on their cognitive load, i.e., total amount of information in their stream. The number of messages in a user's stream is, on average, proportional to the number of friends he or she follows; therefore, we divide users into subpopulations based on the number of friends they follow. A dramatically different picture of exposure response emerges. Now, the response of users within each population increases monotonically with the number of exposures, similar to the ICM. However, unlike ICM, the response of better connected users is suppressed, due to the greater demands placed on their limited attention. The aggregated exposure response in Fig. 3b appears to saturate, because the better connected, and less responsive, users contribute to the right-hand portion of the exposure curve.

This result gives us a better picture of what is going on. Unlike the spread of a virus, which is boosted by hubs, or highly connected people, who create multiple opportunities for the virus to spread, information cascades are suppressed by such users. A cascade stops when it reaches such hubs, because they are less likely to see the message and retweet it, since there are so many other messages competing for their limited attention. Once the response of the highly connected users is encoded into a model of contagion, it leads to smaller cascades.

Discussion

Access to large data sets containing traces of social interactions has created new opportunities to study social behavior. One of the main challenges in analyzing such data is its heterogeneity. People vary greatly in their abilities and motivations, and aggregating over all individuals can sometime lead to erroneous conclusions. This effect, known as "heterogeneity's ruses" (Vaupel and Yashin 1985), was demonstrated above in how people respond to exposures to information in social media. When averaged over all users, it may appear that the more times an individual is exposed to information, the less likely he or she is to spread it. However, when we divide people into more homogeneous populations based on the number of friends they follow, exposure response changes qualitatively. Now, individual response within each population increases monotonically: the more times a user sees information, the more likely he or she is to spread it. However, users with more friends are overall less sensitive than users with few friends. The revised exposure response explains a puzzling observation with which we started this chapter: information in social media does not spread very far. It appears that decreased sensitivity to exposure of highly connected people inhibits social contagion and prevents information from spreading.

The challenge of analysis is to segment the data appropriately. In our analysis, we divided people into classes based on their cognitive load, or the volume of information in their stream. This decision was motivated by our discovery of the role that limited attention plays in the spread of information in social media. Users appear to expend finite effort or time on discovering content. Since users with many active friends have many more messages in their stream to process than users with few

friends, the well connected users are less likely to discover, and spread, any specific message. For other problems, other segmentations of data may be desirable. As the amount of social data increases, finer segmentations of data into more homogeneous populations will be statistically feasible, leading to finer-grained models of human behavior.

References

Adamic LA, Adar E (2005) How to search a social network. Soc Netw 27(3):187–203

Anagnostopoulos A, Kumar R, Mahdian M (2008) Influence and correlation in social networks. In: Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining, Las Vegas. ACM, New York, pp 7–15

Castellano C, Fortunato S, Loreto V (2009) Statistical physics of social dynamics. Rev Mod Phys 81(2):591–646

Clauset A, Shalizi CR, Newman MEJ (2009) Power-law distributions in empirical data. SIAM Rev 51(4):661+

Goel S, Watts DJ, Goldstein DG (2012) The structure of online diffusion networks. In: Proceedings of the 13th ACM conference on electronic commerce (EC 2012), Valencia

Gruhl D, Liben-nowell D (2004) Information diffusion through blogspace. In: Proceedings of the international world wide web conference (WWW), Geneva, pp 491–501

Hethcote HW (2000) The mathematics of infectious diseases. SIAM Rev 42(4):599–653

Hodas N, Lerman K (2012) How limited visibility and divided attention constrain social contagion. In: ASE/IEEE international conference on social computing, Amsterdam

Hogg T, Szabo G (2009) Diversity of user activity and content quality in online communities. In: Proceedings of international conference on weblogs and social media (ICWSM), San Jose

Kahneman D (1973) Attention and effort. Prentice Hall, Englewood Cliffs

Kempe D, Kleinberg J, Tardos É (2003) Maximizing the spread of influence through a social network. In: KDD '03: proceedings of 9th international conference on knowledge discovery and data mining, Washington DC, pp 137–146

Lerman K (2007) Social information processing in social news aggregation. IEEE Intern Comput Spl Issue Soc Search 11(6):16–28

Lerman K, Ghosh R, Surachawala T (2012) Social contagion: an empirical study of information spread on digg and twitter follower graphs. arXiv:1202.3162

Lotan G, Graeff E, Ananny M, Gaffney D, Pearce I, Boyd D (2011) The revolutions were tweeted: information flows during the 2011 Tunisian and Egyptian revolutions. Int J Commun 5:1375–1405

Moreno Y, Pastor-Satorras R, Vespignani A (2002) Epidemic outbreaks in complex heterogeneous networks. Eur Phys J B Condens Matter Complex Syst 26(4):521–529

Newman MEJ (2022) Spread of epidemic disease on networks. Phys Rev E 66(1):016128+

Pickard G, Pan W, Rahwan I, Cebrian M, Crane R, Madan A, Pentland A (2011) Time-critical social mobilization. Science 334(6055):509–512

Romero DM, Meeder B, Kleinberg J (2011) Differences in the mechanics of information DiffusionAcross topics: idioms, political hashtags, and complexcontagion on twitter. In: Proceedings of world wide web conference, Lyon

Salganik MJ, Dodds PS, Watts DJ (2006) Experimental study of inequality and unpredictability in an artificial cultural market. Science 311:854

Satorras RP, Vespignani A (2001) Epidemic spreading in scale-free networks. Phys Rev Lett 86(14):3200–3203

Ver Steeg G, Ghosh R, Lerman K (2011) What stops social epidemics? In: Proceedings of 5th international conference on weblogs and social media, Barcelona

Vaupel JW, Yashin AI (1985) Heterogeneity's ruses: some surprising effects of selection on population dynamics. Am Stat 39(3):176–185

Wang Y, Chakrabarti D, Wang C, Faloutsos C (2003) Epidemic spreading in real networks: an eigenvalue viewpoint. In: IEEE symposium on reliable distributed systems, Florence 0:25+

Wilkinson DM (2008) Strong regularities in online peer production. In: EC'08: Proceedings of 9th conference on electronic commerce, Chicago. ACM, New York, pp 302–309

Wu F, Huberman BA (2007) Novelty and collective attention. Proc Natl Acad Sci 104(45):17599–17601

Wu F, Huberman B, Adamic L, Tyler J (2004) Information flow in social groups. Phys A 337(1): 327–335