Digging in the Digg Social News Website

Siyu Tang, Norbert Blenn, Christian Doerr, and Piet Van Mieghem

***Abstract*—The rise of social media aggregating websites provides platforms where users can actively publish, evaluate, and disseminate content in a collaborative way. In this paper, we present a large-scale empirical study about "Digg.com", one of the biggest social media aggregating websites. Our analysis is based on crawls of 1.5 million users and 10 million published stories on Digg. We study the distinct network structure, the collaborative user characteristics, and the content dissemination process on Digg. We empirically illustrate that friendship relations are used effectively in disseminating half of the content, although there exists a high overlap between the interests of friends. A successful content dissemination process can also be performed by random users who are browsing and digging stories. Since 88% of the published content on Digg is defined as news, it is important for the content to obtain sufficient votes in a short period of time before becoming obsolete. Finally, we show that the synchronization of users' activities in time is the key to a successful content dissemination process. The dynamics between users' voting activities consequently decrease the efficiency of friendship relations during content dissemination. The results presented in this paper define basic observations and measurements to understand the underlying mechanism of disseminating content in current online social news aggregators. These findings are helpful to understand the influence of service interfaces and user behaviors on content dissemination.**

***Index Terms*—Content dissemination, friendship relations, social media website, user characteristics.**

I. Introduction

SOCIAL *media aggregator websites* (in short, social aggregators) such as Digg, Reddit, Delicious, and Slashdot are emerging specialized forms of online social networks (OSNs)[1] and begin to shift the way people search for and consume information on the Internet. By incorporating a variety of social features, social media websites allow users to publish, discover, and promote the most interesting content without a group of website editors.

Manuscript received August 31, 2010; revised January 18, 2011 and May 12, 2011; accepted May 26, 2011. Date of publication June 16, 2011; date of current version September 16, 2011. The work was supported in part by TRANS (http://www.trans-research.nl). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Daniel Gatica-Perez.

The authors are with the Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, 2600 GA Delft, The Netherlands (e-mail: S.Tang@tudelft.nl; N.Blenn@tudelft.nl; C.Doerr@tudelft.nl; P.F.A.VanMieghem@tudelft.nl).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2011.2159706

[1]News portals (e.g., Digg, Reddit, Delicious, etc.) are frequently referred to as social media websites. However, as these sites incorporate and heavily make use of social features, such as the ability to form friendships, follow one's activities, or facilitate one-on-one interaction, these services have become in principle specialized types of OSNs.

In general, social media websites are characterized by the following features: 1) users submit content published from different resources and "aggregate" information on a single website; 2) asocial media website facilitates a rating/recommendation system so that users can vote for and recommend the most interesting content to others in a collaborative way; 3) users can specify their social profiles, designate friends, and connect with people with similar interests; 4) users can discover information on the social media websites by actively tracking their friends' activities, or following different interfaces provided by the websites, e.g., the recommendation engine, popular/unpopular content, different categories, and web widgets incorporated on external websites.

Although discovering information from friends is still an appealing feature, users are encouraged to actively access content via various interfaces provided by these social media websites. The different means for users to explore and disseminate information may directly lead to distinguishable content dissemination patterns in such networks. The process of publishing new content and obtaining users' votes is called *content dissemination*[2] in this paper.

In this paper, we present a large-scale empirical study of Digg.com, a popular social media aggregator. Our objective is to investigate the distinct usage behaviors of Digg users, the patterns of information spread, as well as the impact of friendship relations on content dissemination on the Digg platform.[3] A detailed understanding of these processes has many applications, for example in viral marketing [27], online campaigns [9] and targeted advertisements [38], or innovation adoption.

In the following, we describe three major contributions:

- First, traditional crawling techniques, such as the breadth-first search (BFS), explore users following their friendship relations, but fail to discover users who are actively visiting the network without designating any friends. To avoid overlooking users without friends, we initiate a crawling process by a simultaneous exploration of the network from multiple perspectives. The crawling methodology presented in this paper allows us to capture the most valid and comprehensive information about friendships, activities of users, and the published content in the Digg network.

[2]Dissemination, propagation, and spread are interchangeable terminologies in this paper. The same applies for content, information, and story.

[3]It should be noted that the semantics of friendship on Digg (i.e., "following" a user as described in Section IV) differs from the friendship relationships found in Facebook, Orkut, or LinkedIn where links are formed based on real-world acquaintance or business relationships [6]). Moreover, the main function of Digg.com (i.e., to share information) also varies from other social networking sites (i.e., to network with friends). Hence, the results presented in this paper are applicable to social media websites that have similar functionality as Digg.com. To what extent these findings can be extended towards other types of OSNs needs more investigation.

- Second, by using the collected traces, we study the Digg friendship network and compare its topological properties with different OSNs. We also evaluate the distinct interests of friend pairs in the Digg network and provide empirical evidence to substantiate the common hypothesis of "friends are sharing similar interests in OSNs".
- Third, we analyze the collective information dissemination patterns on Digg. We show empirically that friendship relations can only effectively disseminate Digg content in half of the cases. The friend pairs who are active during content propagation only account for 2% of the total amount of friend pairs in the Digg network. In addition, we find that content does not propagate widely over Digg. On average, information is disseminated no further than 3.9 hops away from the original submitter of the story, whereas the average path length of the Digg friendship network is 5.6 hops. Hence, we question the applicability of viral marketing on social media websites that are similar to Digg. We also show that the published content on Digg becomes "hot" in a short period of time and saturates very quickly after becoming popular. Most importantly, synchronization of the users' collaborative voting activities is the key to successful online advertising or marketing on social media websites that accommodates highly transient content with short life duration.

The remainder of this paper is organized as follows. Section II discusses related work on OSNs and social media aggregators. Section III introduces Digg.com and our methodology of crawling the Digg dataset. In Section IV, we study the topological properties of the Digg friendship network and the shared interests between Digg friends. Section V describes the user characteristics and content dissemination patterns on Digg. In Section VI, we discuss the efficiency of the Digg friendship network in the spread of information and highlight the importance of synchronized user activities when disseminating content. Section VII summarizes our findings and outlines future research.

II. Related Work

Early studies on information dissemination can be traced back to the 1950s. Researches at the early stage mainly focused on real-world social systems: for example, Rogers studied the processes and theories to diffuse ideas, practices, and innovations in social systems [32]; Katz and Lazarfeld investigated the origin and spread of influence through social communities [16]. With the thriving of OSNs in recent years, much more research efforts were devoted to the field of online social communities, as they provide an easily accessible and large-scale data source which reflects the collaborative behaviors of millions of users as well as the way that information is disseminated among them. In the following, we briefly review related work performed with different OSNs and with Digg.com.

A. Online Social Networks in General

Mislove *et al.* [25] studied the topological properties of four OSNs: Flickr, YouTube, LiveJournal, and Orkut. They obtained the data by crawling publicly accessible information on these networking sites. In [25], it was found that the studied OSNs display both power-law and small-world properties, e.g., tightly connected clusters of nodes, high levels of link symmetry, and a positively correlated node degree. Mislove *et al.* also discussed the impact of the observed network structure on the spread of information and the implication for the design of dissemination/search algorithms for OSNs. Kossinets *et al.* [17] studied e-mail communication within a university over a two-year period. They analyzed the underlying social network and identified the information "backbone", on which information has the potential to flow the quickest.

Leskovec *et al.* [21] presented an extensive analysis about the communication behaviors and characteristics of the Microsoft Messenger instant-messaging (IM) users. They examined the communication patterns of 30 billion conversations among 240 million people, and found that people with similar characteristics (e.g., age, language, and geographical location) tend to communicate more. Liben-Nowell *et al.* [22] analyzed the geographical location of LiveJournal users and found a strong correlation between friendship and their geographic proximity. Benevenuto *et al.* [5] examined user activities of Orkut, MySpace, Hi5, and LinkedIn. A clickstream model was presented to characterize user interaction between OSNs friends, as well as how users switch from one activity to the next in such an OSN (e.g., search for profiles, browse friends' pages, send messages to friends).

Another popular topic about OSN focuses on the process of content dissemination. For instance, Cha *et al.* [8] studied photo propagation patterns in the Flickr OSN and the impact of social relations on the propagation of photos. The results in [8] suggest that friendship relations play an important role during information spread as over 50% of users find their favorite pictures from their friends in the social network. It was shown that there are different photo propagation patterns in Flickr and photo popularity may increase steadily over years. A similar study was performed with YouTube and Daum (a Korean OSN) in [7], where the evolution of video popularity evolution was discussed. Contrary to the findings on Flickr [8], video popularity of YouTube and Daum is mostly determined at the early stage after a video content has been submitted.

B. Studies on Digg.com in Particular

Previous research efforts on Digg have aimed to understand the dynamics of information spread [19], the interaction between a user's influence and information dissemination [18], and the methodology to predict content popularity [20], [33].

Lerman *et al.* [19] showed that the social relations of users play a crucial role in the spread of information on both Digg and Twitter. Content in Twitter spreads continuously as the story ages, whereas on Digg, stories only initially spread quickly through the network. The same observation was found in [18]. Szabo *et al.* [33] analyzed and compared the popularity of Digg content with YouTube videos. As shown in [33], YouTube videos keep attracting views throughout their lifetimes, whereas Digg stories saturate quickly. The influence of a user's relationships is not effective once Digg content has been exposed to a wide audience, although it is important in the early stages when content is only exposed to a small number of users. In [20], the evolution of the popularity of a single story was characterized. In their proposed model, a set of variables is introduced to specify the influence of individual behavior and the effect of user interfaces.

C. Discussion on Previous Research on Digg.com

A fundamental assumption of previous research is that information is primarily disseminated along social links and that the underlying social network is the key to the spread of content in online social communities [6], [8], [25]. The same conclusion was also made for Digg [18], [19], [33].

However, we believe that previous reports on information dissemination do not accurately describe the global processes on Digg.com. For instance, results in [18] and [19] are based on a limited number of stories being collected in a short period of time (around a week). There are in total 2858 stories examined in [18] and 3553 stories analyzed in [19]. Besides, none of the above reports aimed to carry out a comprehensive crawling process to obtain related information on the large-scale Digg website. Although a larger dataset (including 29 million diggs by 560 000 users) is evaluated in [33], it only accounts for less than half of the entire Digg network (as will be shown in Section III). Moreover, previous analyses on Digg focused on the dissemination process from the source via friends, while ignoring the influence of users who do not establish any social connections.

In this paper, we carry out an in-depth analysis on Digg.com from the perspectives of network structure, user characteristics, and content dissemination patterns. Some of our findings, such as the characteristics of user similarity, the submission and digging patterns of users, as well as the impact of social relations on the spread of information, have not been thoroughly investigated before; hence, they may provide important insights in understanding social media websites that are similar to Digg.

III. Digg.com and Data Collection

The social news website Digg.com[4] is a content discovery and sharing application launched in 2004. According to the traffic statistics provided by Alexa.com in May 2010, Digg is rated as the 117th most popular website globally, and as 52nd in the United States. The Digg users can submit a uniform resource locator (URL) of a video, an image, or news that is published elsewhere on the web. By using the voting system on Digg, a user can *digg* a story if he has a positive attitude towards it.

Within Digg, one can create friendship connection to others. A user can either be a *fan* or a *mutual friend* to another person, which is similar as the *follower* relation on Twitter[5]. Via a so-called *friends' activity* interface, each user maintains a *friends list*, i.e., a list of friends that he has designated. If a user A designates user B as a friend in his friends list, user A is specified as a *fan* in user B's list. If B also reciprocates A as a friend, both users are marked as mutual friends. After designating new friends in his list, a user can follow the recent activities of his friends, e.g., stories his friends submitted, commented, or dugg. Similarly, once digging (equivalent as voting for) a story, a user is implicitly disseminating and recommending the content to his fans.

New submissions on Digg are displayed on the *upcoming* section of the website. Stories that are considered to be interesting are selected from the upcoming section and thereafter appear in the *popular* section which is the default page shown to a user entering the Digg website. Promoting an upcoming story is managed by a secret algorithm developed by Digg. The algorithm considers the number of diggs, the diversity of users who are digging the story, the time when the story was submitted, and the topic of the story as the major factors of the promotion.[6] Further details about the algorithm, however, are not provided by Digg. At the moment of our work, there are approximately 15 000 to 26 000 stories being submitted daily, out of which, around 150 stories are promoted to be popular per day. According to our Digg dataset, 88% of the published content on Digg are news, 8.0% are videos, and 4.0% are images.

To explore Digg.com, a user has two options. He can go to the upcoming or the popular section of the website. Stories on each section are organized 1) by type, i.e., news, videos, and images; 2) by topic, i.e., eight major topics of technology, world & business, science, gaming, lifestyle, entertainment, sports, and offbeat; 3) or by a recommendation system, i.e., stories that are ranked with respect to their votes (diggs). Apart from the above interfaces, a Digg user may also discover stories via the friends' activity interface mentioned above. Once logged in, users can keep track of their friends' activities and therefore disseminate information along the friendship relations.

While most social network traces are crawled using friendship relations [1], [25], the Digg dataset was obtained by a simultaneous exploration of the network from four different perspectives as shown in Fig. 1. By using the Digg Application Programming Interface (API), we are able to explore the four perspectives (from bottom to top in Fig. 1) during data collection:

- **Site perspective**: The Digg website lists all popular and upcoming stories under different topic areas. Every hour, we retrieve all popular stories (for all topics) that are listed on Digg. Every four hours, all upcoming stories (for all topics) are collected. All discovered stories are added to an "all-known story" list maintained by us.
- **Story perspective**: For each of the stories that has been retrieved, a complete list of all activities performed by different users (who dugg on the story) is collected. Any user who is discovered will be added to the "all-known user" list for future exploration.
- **User perspective**: For each user discovered within the Digg OSN, the list of their activities, such as submitting and digging on stories, is retrieved. Occasionally, a previously unknown story is discovered (this is typically the case for older stories before we started the collection). For such a story, the entire (digging) activities of users are retrieved for that story.
- **Social network perspective**: Each registered user can make friends with other Digg users. In the crawling process, a list of friends is retrieved for every user. If a friend is a previously unknown user, this user is added to the data discovery process, and a list of all his friends and his public user profile information are retrieved.

The above procedure is continued until no new user and story can be found and periodically repeated afterwards to discover

[4]On August 25, 2010, the Digg platform was updated to Digg version 4. The work presented here is performed on Digg version 3.

[5]Twitter: http://twitter.com.

[6]Content promotion refers to the procedure of obtaining votes from the publication of a story until it is displayed on the front page of the popular section. While a content dissemination process is defined as the growth of diggs for a story during its entire life time on Digg.

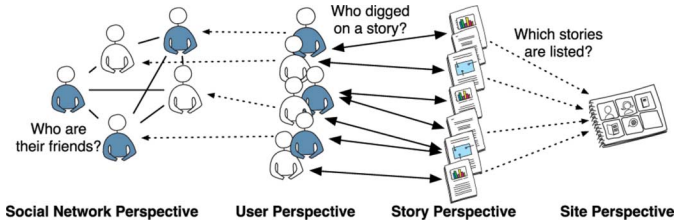

Fig. 1. Different components of the Digg crawling process.

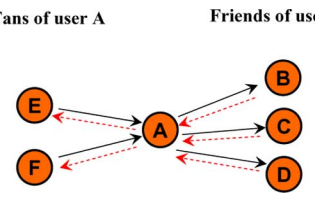

Fig. 2. Illustration of information dissemination in the Digg friendship network. The solid arrows represent the friend or fan relationships. The dotted arrows refer to information dissemination along the friendship links.

new user, story, activity, and friendship relation. By using the above crawling methodology, we are able to collect almost the entire information about friendships and activities of users and the published content on Digg. Although the Digg network was officially founded on December, 2004, the feature of the friends list was not released until July 2005, when the second version of Digg was launched. Therefore, in our analysis, the collected friendship information starts from July 2005 until May 2009. Our Digg dataset has a volume of more than 600 GB (gigabytes), containing the related information about 1.5 million registered users and 10 million published stories in the Digg OSN.

IV. Digg Friends and the Friendship Network

The traditional way to study complex networks such as social networks, biological networks, and the internet is by examining their topological properties. Since friendships are assumed to be critical during content dissemination in OSNs, studying the topological properties of OSNs is useful to understand the way that information is disseminated between users. By considering the 1 527 818 registered Digg users in our dataset as nodes and their follow-relations (fans or mutual friends connections) as links, we construct a directed Digg *friendship network*[7] G_F. A bi-directional link is called a *symmetric link*. Otherwise, the link is referred to as being *asymmetric*. In G_F, the outgoing degree D_{out} defines the number of friends that a random user has, and the incoming degree D_{in} indicates the number of fans that a user has.

As discussed in Section III, users can keep track of their friends' activities and further disseminate content along their friendship connections. Hence, content dissemination in the Digg friendship network is initiated in the reversed direction along the friendship links. In Fig. 2, we illustrate the friend and fan relationships on Digg (the solid arrows), as well as the way that information is disseminated via friendship links (the dotted arrows). For example, user A has three friends and two fans; see Fig. 2. User A may discover (and digg) stories recommended by B, C, and D. A is also a potential source from whom user E and F may acquire information. In short, a user disseminates content to his fans in the opposite direction along the incoming links. In the following, we examine the topological properties of the Digg network and characterize the shared interests between Digg friends.

[7]Notice that the friendship on Digg differs from social connections between individuals in the real word—it is in fact a following relationship between Digg users. We advise readers to differentiate a friendship on Digg as in other OSNs such as MySpace or Facebook.

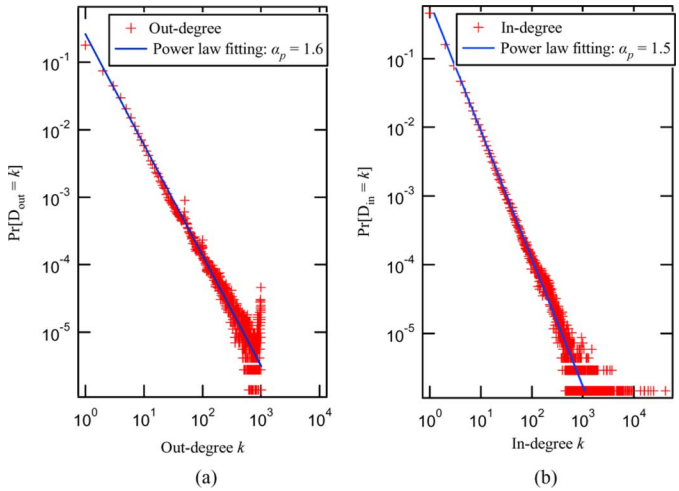

Fig. 3. (a) Pdf of the out-degree of the giant component. A cut-off out-degree of 1000 is set on Digg. (b: Pdf of the in-degree of the giant component. Both curves are plotted on log-log scale and best fitted with the power law distribution. The goodness-of-fit, i.e., $\rho = 0.96$ and $\rho = 0.97$, of the out-degree and in-degree distribution suggest high fitting quality.

A. Network Connectivity and Node Degree

Our analysis shows that there exists a *giant component*[8] in the Digg friendship network. The presence of the giant component is critical to the connectivity and the dissemination of information on Digg. The giant component in our Digg data collection consists of 685 719 nodes (which account for approximately 44% of our entire Digg dataset) and 6 736 174 links. Around half of the users in the giant component have no outgoing links. The average shortest path length and the network diameter[9] of the Digg friendship network is 5.6 hops and 29 hops, respectively. The remarkably short path length and diameter of the giant component is likely to impact the content dissemination process on Digg, as information is not expected to spread far away from the source. In Fig. 3(a) and (b), we plot the probability density function (pdf) of the out-degree and in-degree of the giant component, respectively. On a log-log scale, both $\Pr[D_{out} = k]$ and $\Pr[D_{in} = k]$ exhibit straight lines, conforming to the power law distribution.[10] Notice that Digg artificially caps a user's maximal out-degree to 1000, the distribution of very high out-degrees (close to $D_{out} = 1000$) is skewed in Fig. 3(a). The exponent of $\Pr[D_{out} = k]$ is found to be $\alpha_p = 1.6$, which is slightly higher than the exponent of the in-degree (i.e., $\alpha_p = 1.5$).

[8]We define the giant component as the largest connected component in a network.

[9]The average shortest path length is defined as the average number of hops along the shortest paths for all possible node pairs in a network. The network diameter is the maximal number of hops along the shortest paths.

[10]The power law distribution is defined as $\Pr[X \leq x] = g(x)x^{-\alpha+1}$, where $g(x)$ is a slowly varying function.

Apart from the single giant component, the Digg friendship network also consists of a significant number of connected components and disconnected nodes. There are 31 513 nodes forming 13 270 small and distinct connected components. These *connected components* are not connected to the giant component. The maximum number of nodes in these components is 77, and the smallest component only consists of 2 nodes. The remaining 810 586 *disconnected nodes*, accounting for about half of all Digg users, are not connected to any other nodes in the network. Our observation about the node degree distribution and connectivity on Digg suggest two important findings.

First, the Digg node degree distribution is consistent with previous reports on Flickr, YouTube, LiveJournal, and Orkut (see [25]), in the sense that the power law exponents of these OSNs are all smaller than 2 (between 1.5 and 2). The node degree distribution of OSNs implies their fundamental difference compared with the structure of other complex networks, e.g., real-world social network, World Wide Web (www), and power grid network, in which the power law exponents are between 2.1 and 4 [4]. The observed power law distributions with an exponent smaller than 2 indicates that, when $N \rightarrow \infty$, all moments including the mean tend to ∞. Hence, the degree of the nodes exhibit high variations.

Second, most large OSN traces (e.g., in [8] and [24]) are crawled by using the breadth-first search (BFS) technique following friendship links. Although it is shown in [25] that the number of missing users from the giant component by using BFS tend to be small in number, the amount of disconnected nodes are naturally excluded from the collected dataset. Moreover, analysis of users that are not in the giant component are usually ignored, and their impact on OSNs has not been well studied. With the Digg dataset collected in this paper, we are able to analyze the characteristics of the disconnected users and their digging behaviors, which will be discussed in more detail in Section V-AII.

B. Degree Correlation

The *assortativity* measures the degree correlation of connected nodes in a graph. The *assortativity coefficient* ρ is essentially the Pearson correlation coefficient [34, p. 30] between two random variables X and Y, and lies in the range of $-1 \leq \rho \leq 1$. A positive assortativity coefficient (close to 1) indicates an *assortative* network in which nodes are likely to connect to others of similar degree, while a negative assortativity coefficient (close to -1) refers to a *disassortative* network where high-degree nodes tend to connect to low-degree nodes. In theory, disassortativity favors good connectivity in a network while nodes in an assortative network exchange information with those that reward them equally [36].

According to Newman's definition [29], the assortativity in directed networks measures the tendency of a node i to connect with other nodes that have incoming degrees similar to node i's outgoing degree. Here, we measure the assortativity as the in-degree (or out-degree) correlation between two connected nodes, because information is always disseminated in the opposite direction of incoming links. Therefore, the revised definition of assortativity is useful when examining the impact of network structure on content dissemination.

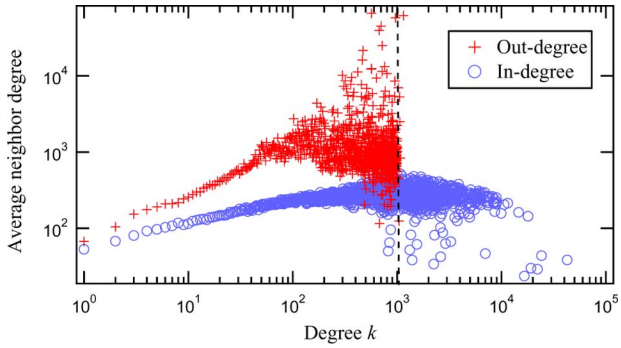

Fig. 4. Log-log plot of the out-degree (in-degree) versus the average out-degree (in-degree) of neighbors. The vertical dotted line shows Digg's out-degree limitation of 1000.

Our analysis shows that there exist no significant relations regarding the in-degree between two Digg friends ($\rho_{in} = -0.03$). In terms of content dissemination, an assortative network is more favorable since individuals are tightly connected to others of high in-degree, with the low in-degree nodes on the edge of the network [29]. Hence, information can spread quickly and effectively from a user to his many neighbors and further within these groups of users. However, such a property is not observed in the Digg network.

The out-degree assortativity, on the other hand, provides insight on the correlation between a pair of friends regarding the number of sources from which they can discover content. The derived assortativity coefficient of $\rho_{out} = 0.05$ indicates that users who make many friends might not connect to those are similar (i.e., users who also keep many friends). The content discovery process can be performed more efficiently and widely in an assortative network (regarding the out-degree) since a search query reaches more neighbors. Apparently, the Digg network does not have such a property either.

To investigate the assortativity in more detail, we plotted the average neighbor out-degree (in-degree) distribution of nodes with out-degree (in-degree) k in the Digg friendship network in Fig. 4. An increasing average neighbor degree distribution indicates a tendency of higher-degree nodes to connect to other high-degree nodes, whereas a decreasing distribution represents the opposite trend. Fig. 4 illustrates an increasing trend for lower degree nodes (e.g., nodes with out-degree smaller than 100) to connect to higher degree nodes. However, as the out-degree (in-degree) grows, we see drastic fluctuations. Thus, there is a nonlinear relationship between the degrees of highly connected nodes and the degrees of their neighbors.

C. Link Symmetry

In OSNs, bi-directional links reveal a reciprocal relation between users. Link reciprocity in OSNs is fundamental, because the presence of mutual links can accelerate the information propagation process and reduce the diameter during content discovery [12].

Contrary to previously studied OSNs which exhibit a high level of symmetry (e.g., 62% of links in Flickr, 74% of links in LiveJournal, and 79% of links in YouTube are found to be bi-directional [25]), the link symmetry in the Digg network is much lower (38% on average) and varies with respect to the degree of

TABLE I
FRACTION OF SYMMETRIC LINKS ON DIGG

User Group	Number of Users	Symmetric Links (Fraction)
$0 < D_{out} < 10$	282536	53%
$10 \leq D_{out} < 100$	49416	42%
$100 \leq D_{out} < 1000$	13993	39%
$D_{out} = 1000$	111	31%

nodes. As shown in Table I, users that are connected to a small group of friends (e.g., $0 < D_{out} < 10$) are more likely to designate each other as mutual friends (53% of their friendship links are symmetric). Creating many connections does not increase the probability of being accepted as a friend: only 31% of the friendship links are bi-directional for users with $D_{out} = 1000$.

The lower friendship-reciprocity on Digg is due to the lack of incentives for establishing mutual friendships. In fact, the Digg website uses an "asymmetric model" by intentionally distinguishing between "fans" and "mutual friends". In contrast, previously considered OSNs (e.g., Flickr, LiveJournal, and YouTube) only allow for mutual friendships to be established, leading to the high symmetric link ratios observed in these networks. The asymmetric model used on Digg is consistent with sociological studies, which predicted the necessity of specifying the diversity of possible relationships amongst users and distinguishing strong (mutual friends) and less strong (the fans) friendship relations [11], [30]. However, the low level of link symmetry points to a less favorable situation for content discovery and sharing between friends.

D. Shared Interests Between Friends

According to sociological theory, friends in OSNs tend to have common interests and tastes [3], [31]. Hence, within Digg, it is also assumed that users browse stories dugg by other members and establish friendship relations if they share the same interests. In the following, we characterize users' interests on Digg and compare the taste similarity between friends.

Recall that all stories on Digg are classified into eight major topics. The number of stories that a user dugg under each individual topic reflects his general interests on that specific topic area. We denote $X_k (1 \leq k \leq 8)$ the number of stories a user has dugg[11] under topic k. The elements of the set $\{X_{(1)}, X_{(2)}, \ldots, X_{(8)}\}$ are the ranked random variables of $\{X_k\}_{1 \leq k \leq 8}$, if $X_{(1)} = \max_{1 \leq k \leq 8} X_k$, and $X_{(8)} = \min_{1 \leq k \leq 8} X_k$. The sum $S = \sum_{k=1}^{8} X_k$ is the total number of stories that a user has dugg. Consequently, $R_k = X_{(k)}/S$ defines the fraction of stories a user has dugg under his kth favorite topic over the entire amount of stories dugg by him.

Fig. 5 depicts $E[R_k|V_n]$, the average value of R_k provided that a user is interested in $n (1 \leq n \leq 8)$ topic areas (denoted by the event V_n). The dotted (vertical) lines in Fig. 5 represent the group of users that are interested in n topic areas. We see

[11]The index of k from 1 to 8 corresponds to the topic of technology, world & business, science, gaming, lifestyle, entertainment, sports, and offbeat, respectively.

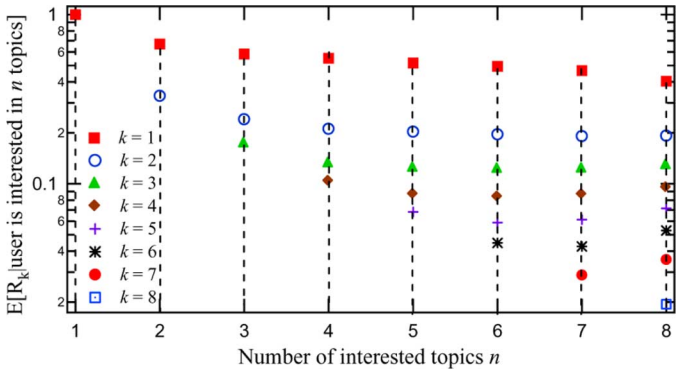

Fig. 5. Rankings of user interests under individual topics. The horizontal axis represents the number of topic areas in which users are interested. The vertical axis shows the ranking of users' kth favorite topic. It is plotted on logarithmic scale for easier reading.

far more activity under a user's preferred topic as opposed to his less favored topics. For example, if users read, digg, and are interested in two topics, on average, the number of stories under their favorite topic accounts for 67% of the total stories that they dugg (see the second vertical dotted lines). Only 33% of the stories fall into their least favorite topic. Users who are interested in all eight topics digg approximately 40% and 19% of the stories under their two favorite story categories and the number of stories dugg under individual topics decreases logarithmically. In general, users digg at least twice as many stories under their favorite topic as under the second favorite one.

The rankings of user interests provide a direct way of measuring the similarity between users' tastes when making friends. When comparing the interests between two users and their ranking of topics, we define the *similarity hop*, which can be used to reflect the distance of a user's favorite topic with respect to the kth favorite topic of his friends. We denote by $T_{(k)}$ the kth topic after ranking. For a friend pair i and j, we obtain two set of lists of $\{t_{i(1)}, t_{i(2)}, \ldots, t_{i(8)}\}$ and $\{t_{j(1)}, t_{j(2)}, \ldots, t_{j(8)}\}$, in which $t_{i(k)}$ and $t_{j(k)}$ are the names of the kth favorite topic of user i and user j, respectively. Since $t_{i(1)}$ is the most favorite topic of user i, we compare $t_{i(1)}$ with $t_{j(k)} \in \{1 \leq k \leq 8\}$ of user j. The similarity hop, defined in (1), measures how similar two friends regard their tastes:

$$h_{ij} = (k-1)1_{\{t_{i(1)} = t_{j(k)}\}} \tag{1}$$

in which the indicator function, $1_{\{x\}}$, is defined as 1 if the condition x is satisfied; else, it is zero. The similarity hop h_{ij} ranges between $0 \leq h_{ij} \leq 7$. A zero hopcount means that two users have identical interests. A small similarity hopcount, say $h_{ij} = 1$, indicates high overlapping interests between two friends, while a large hopcount (e.g., $h_{ij} = 7$) suggests that users do not have common interests.

Our analysis shows that the similarity hop between two friends decreases exponentially: 36% of friend pairs have identical interests and the percentage of friend pairs that are one, two, and three hops away are 20%, 15%, and 10%, respectively. On average, the similarity hops between all the friend pairs on Digg is calculated as 1.7, indicating a quite high overlap in user interests. The above analysis provides direct evidence to the common assumption of "friends share similar interests in OSNs".

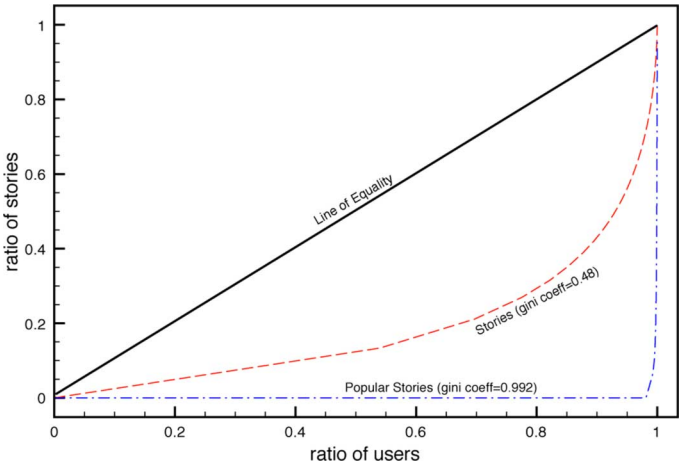

Fig. 6. Stories submission pattern of the 1.5 million Digg users. Out of the 10 million stories submitted to Digg, only 1% of them become popular.

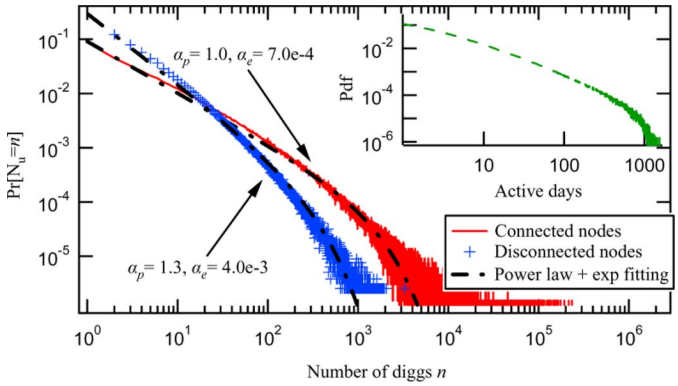

Fig. 7. Pdf of the number of stories dugg by the connected and disconnected users. The best fitting curve is the power law distribution with an exponential cut-off. Inset: Pdf of the number of active days of Digg users.

V. User Characteristics and Content Dissemination on Digg

Users are the driving force to publish, vote, and recommend content on Digg. In this section, we first evaluate the distinct characteristics and digging activities of the Digg users. Afterwards, we study how content is disseminated on Digg.com after their publication.

A. Characteristics of User Activity

Social media websites such as Digg.com allow users to submit their content. Hence, we are motivated to evaluate whether users are actively utilizing this new feature (e.g., submitting and digging stories) on these websites.

1) Content Submission on Digg: Fig. 6 plots the Lorenz curve [23] of the story submission pattern on Digg. While the 10 million stories published on Digg are supposed to be submitted by the 1.5 million users, we see that 80% users are in fact only submitting approximately 20% of the entire Digg content. While it is a quite unbalanced system,[12] the above observation conforms to the Pareto principle, i.e., the "80–20 rule" [15], that has been widely observed in economics and sociology. The inequality of story submission becomes more drastic for popular stories. Only a small group (2%) of users have succeeded in submitting popular stories, whereas the majority of Digg submitters fail to make their content to be successful.

Even though social media websites are considered to be a platform that provide equal opportunity for users to disseminate their stories, the "flavor" of Digg.com is in fact dominated by a minority of people in the social community. The presence of a small group of successful users seemingly suggests that they are the critical users who can effectively disseminate content. However, our analysis shows that these 2% users are not always successful in submitting popular stories. First, there is no correlation between the number of stories submitted by these users and the stories that become popular (the Pearson correlation coefficient is -0.02). Second, the average ratio of submitted popular stories of the 2% successful users over their total number of submissions is 0.23. Third, we do not find users who can continuously repeat their previous successes over time. Hence, delegating the spread of content to the top 2% users does not always guarantee a successful dissemination.

2) Digging Activity of Connected and Disconnected Users: We define an *active* user as the one who dugg at least one story on Digg and denote N_u as the number of stories a user dugg. Out of the 1.5 million users in our Digg dataset, 93% are active users whereas the remaining 7% of users registered on Digg without digging any story. As discussed in Section IV-A, there exists a huge amount of disconnected users, which accounts for approximately half of the entire Digg users in our dataset. These disconnected users may take advantage of different interfaces provided by Digg, and therefore are actively exploring and digging stories. Hence, it is interesting to examine the digging activities of the connected and disconnected users separately.

Fig. 7 presents the pdf of N_u for the connected and disconnected Digg users, respectively. The pdf of N_u for the connected users has a long-tail shape. The majority of users diggs less than ten stories and a few of them digg more than thousands of stories. The disconnected users are also actively digging stories on the Digg website, indicating that users are indeed using various interfaces provided by Digg to access and digg stories. Friendship relation is no longer the only mean for content discovery and digging on Digg. Fig. 7 shows that the pdf of N_u for the disconnected users also exhibits the heavy-tail property. However, the digging behaviors of the connected and disconnected users exhibit a major and quantitative difference—the disconnected users, in general, are digging less stories than the connected users. The maximal number of dugg stories by the disconnected users is approximately one order less than users in the connected components. The number of disconnected users who dugg less than ten stories is higher than the connected ones.

Moreover, on a log-log scale, there is a slightly bending trend at the end of the straight lines in Fig. 7. For both curves of $\Pr[N_u = n]$, the best fit is the power law distribution with an exponential cut-off[13] at the tail. The power law distribution with an exponentially decaying tail has been reported in protein networks [14], e-mail networks [10], actor networks [2], www networks [28], and for video popularity in YouTube [7]. To explain

[12]The further the Lorenz curves are away from the line of equality, the more unequal the system is. A low Gini coefficient implies the tendency towards an equal system—a zero Gini coefficient corresponds to complete equality, and vice versa.

[13]The power law with an exponential cut-off is described by $f(x) = cx^{-\alpha_p}e^{-\alpha_e x}$, where the exponential decay term $e^{-\alpha_e x}$ overwhelms the power law behavior at very large x.

the generating process of such distribution, several models have been proposed (e.g., the aging effect by Amaral *et al.* [2] and the limited web page availability by Mossa *et al.* [28]). As for the Digg network, the bending tail may be attributed to the saturation of users' digging capability, i.e., users cannot discover and digg all stories published on Digg. The inset of Fig. 7 plots the pdf of the number of active days of users on Digg (a user who dugg at least one story is considered to be active on that day). As we can see, a Digg user can stay active on Digg for maximally about 1000 days. In order to digg the entire 10 million published Digg stories, the user needs to digg 10 000 stories per day, which is not possible. Hence, the probability for users to digg more stories decays faster than a power law.

B. Content Dissemination Pattern

Our analysis shows that the *diggcount*, i.e., the of number of diggs received by a story, is highly correlated with the *pageview*[14] of that story. The correlation coefficient is 0.87. Hence, we consider diggcount as a good metric to reflect the popularity of a Digg story.[15] In this section, we examine the diggcount of the 115 163 popular stories in our dataset as well as the way that these stories are disseminated after their publication.

1) Story Promotion Duration: Dissemination of stories on Digg consists of two phases: before and after they are popular. The promotion duration T of a popular story refers to the time between its publication and promotion. In Fig. 8, we plot the distribution $F_T(t)$ of the promotion duration of the collected 115 163 popular stories. As revealed from Fig. 8, the first 24 hours from the publication is critical to the submitted upcoming stories (15 000 to 26 000 stories per day) on Digg: stories need to attract sufficient interests, e.g., a large enough number of diggs, the diversity of the diggers, within 24 hours in order to get promoted. Fig. 8 also reveals that the average promotion duration is approximately 16.3 h. The promotion pattern on Digg differs from other OSNs (e.g., Flickr), where content may grow steadily over years and finally become "hot". We attribute the distinct feature of promoting stories with stringent timeline to the fact that most Digg content (88%) is social news. Thus, the novelty of the news is critical to attract the attention of users.

2) Identifying Diggs From Friends and Non-Friends: We follow the heuristic approach in [8] and [33] to identify diggs from friends and non-friends. An important assumption is made as: If A is the fan of B, A reads and diggs a story after B dugg it, we say that the story is recommended by user B and therefore disseminated via friendship links. Following this assumption, we are able to infer a digg on a story as the digg from a friend (a user discovers and diggs the story via friendship link) and the digg from a non-friend (a user diggs a story without the engagement of the friendship network) and further evaluate the effectiveness of friendship and non-friendship relations during content dissemination. It should be underlined that it is possible for a user and his fans to discover a story through the Digg website interfaces rather than via friendship relations. Our measurement, however, is examining the upper bound of the friendship network during the spread of information, in the case where friendship relation is the only mean to disseminate information.

[14]The pageview is defined as the number of visitors who has clicked and read a story.

[15]A user is only allowed to digg once on a story.

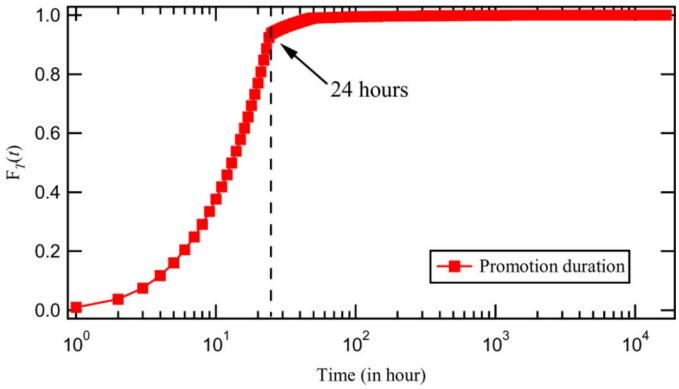

Fig. 8. $F_T(t) = \Pr[T \leq t]$ of the promotion duration of popular stories (in hours). The average promotion duration of popular stories is approximately 16.3 h.

TABLE II
RATIO OF FRIENDS AND NON-FRIENDS OVER THE TOTAL NUMBER OF DIGGERS FOR POPULAR STORIES

	Before popular		After popular	
Average ratio	Friends	Non-friends	Friends	Non-friends
63,484 stories	0.72	0.28	0.25	0.75
51,679 stories	0.23	0.77	0.14	0.86

3) Collaborative Content Dissemination: After identifying a digg from friend and non-friend users (as explained in Section V-BII), we calculate the ratio of friends and non-friends over the total number of diggers before and after a story became popular on Digg. Out of the entire 115 163 popular stories, 55% are predominantly disseminated via social relations (72%) before their promotion, with a minor contribution from non-friends (28%). In the remaining 45% of cases, there is no significant contribution of the friendship relations (23%), and stories are mainly promoted by non-friends (77%) before they became popular. Table II presents the aforementioned ratios of the two types of stories. For both types of stories, the number of diggs from non-friends is significantly larger after stories are promoted. This is because since stories are placed on the front page of the popular section, they become more "visible" and more easily accessible for users who are browsing the websites. Users can still digg stories recommended by their friends, while the influence of the friendship network is marginal once stories are exposed to a vast number of random users who are active on the website.

Fig. 9(a) and (b) presents two typical examples for stories from each type. The number of friends/non-friends, who are digging the story, is plotted as a function of time. We see that story 10471007 [Fig. 9(a)] receives most diggs from friends before its promotion, while story 1083159 [Fig. 9(b)] mainly relies on non-friends for its promotion. For both stories, the number of diggs increases drastically after their promotion. Once stories are placed on the first front page of the popular section, they quickly obtain the attention of many users.

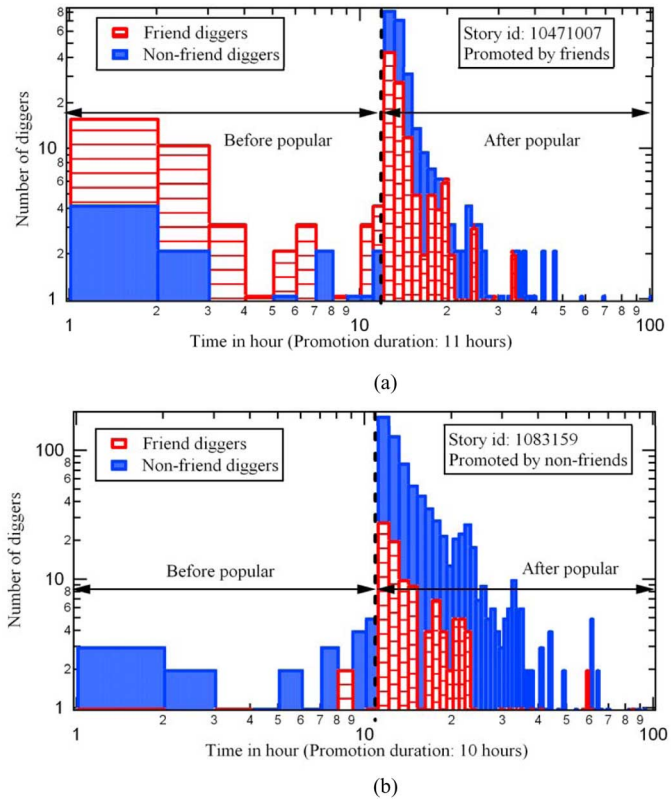

Fig. 9. Dissemination pattern of a story since publication. (a) The story is promoted by friends. (b) The story is promoted by non-friends (log-log scale for easier reading).

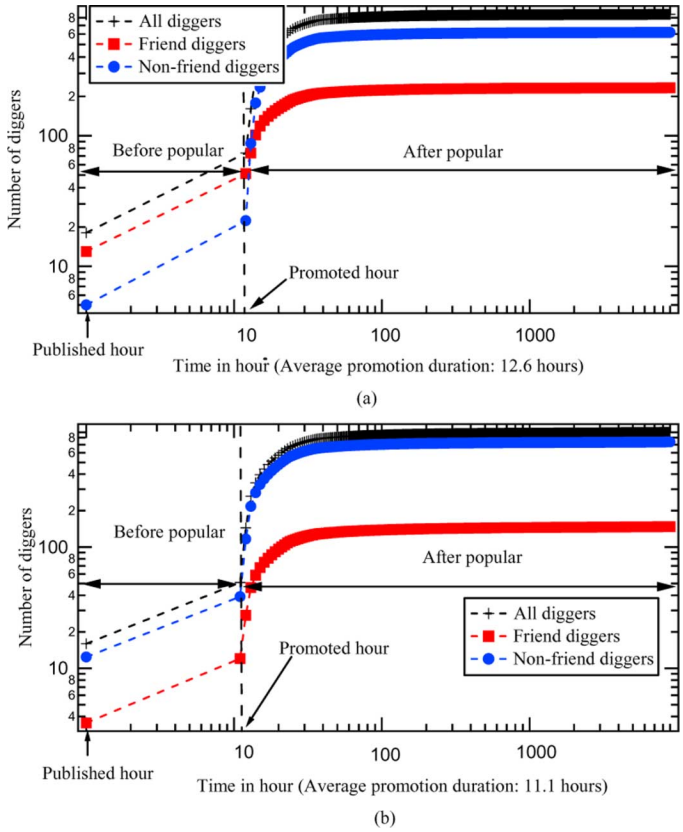

Fig. 10. Aggregated dissemination pattern of popular stories. (a) The 63 484 popular stories promoted by friends (average promotion duration is 12.6 h). (b) The 51 679 stories promoted by non-friends (average promotion duration is 11.1 h) (log-log scale).

In order to illustrate the overall performance of stories in both types, we plot the aggregated dissemination patterns[16] for all stories in each type in Fig. 10(a) and (b). Most diggs are obtained within the first 2 or 3 h after stories are placed on the first front page and the "attractive period" of a story is very short. As time elapses, stories lose their popularity very fast and the number of diggs becomes stable approximately after 40 h since their promotion. The attractiveness of social news is in nature limited by time and becomes obsolete very fast. Besides, stories are shifted gradually from the first front page to the second, third, and so forth. Thus, users may not want to explore stories that are out-of-date.

4) Distribution of Story Popularity: The number of diggcounts of a story reflects its popularity on Digg. The growth of story popularity is a dynamic process and the diggcount of a story varies as a function of time. Hence, we study the pdf of the diggcount for popular stories appearing on different front pages. In particular, we denote $X_m(m \geqslant 1)$ the diggcount of stories on front page m, i.e., the number of diggs a story obtained before it is shifted to the next page. As shown in Fig. 11, the curves (diggcount distribution on the first five front pages) are fitted reasonably well with the lognormal distribution.[17] In fact, as stories are being shifted to subsequent front pages, the lognormal distribution can still describe the story diggcount distribution. After (approximately) the 18th front page, the diggcounts of stories hardly increase.

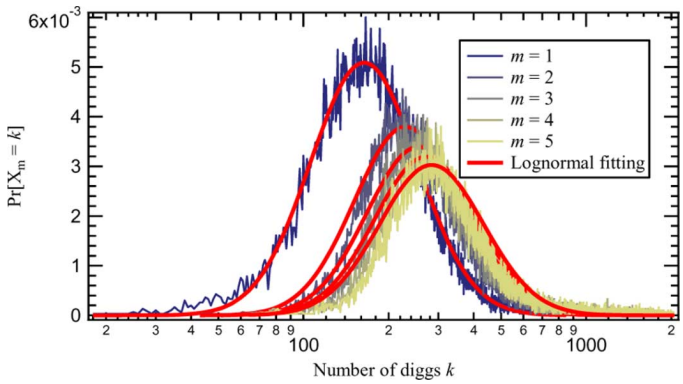

Fig. 11. Pdf of story diggcount on the first five front pages for the popular stories on Digg.

The lognormal distribution of story popularity has been reported in a variety of fields, e.g., collaborative edits per article on Wikipeida [37], growth dynamics of the WWW [13], and the growth of an organism [26]. To explain the process that generates such distribution and predict the diggs received by a story, we refer to the single parameter model developed by Wu and Huberman [39], as well as the model with multiple variables introduced by Lerman and Hogg [20]. While the above

[16]Since stories have different promotion durations, we compute the aggregated number of diggs at the time that a story is published, and the number of diggs of that story when it is promoted to popular. Thus, in Fig. 10, only two time points are plotted before stories are promoted.

[17]The lognormal probability density function with parameters μ and σ is defined as

$$f_{lognormal}(x) = \frac{exp\left[-\frac{(lnx-\mu)^2}{2\sigma^2}\right]}{\sigma x\sqrt{2\pi}}. \quad (2)$$

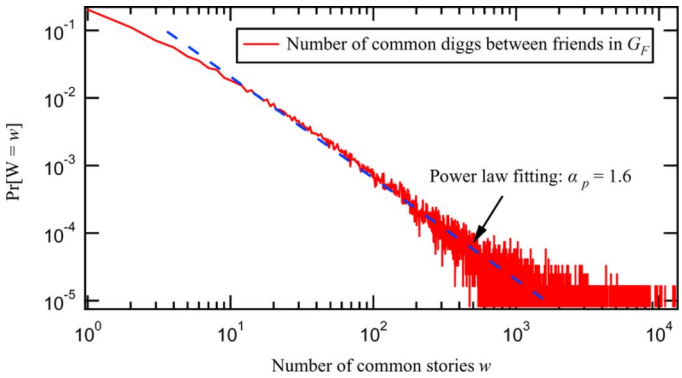

Fig. 12. Pdf of the number of common stories dugg by friends on Digg.

studies give a partial understanding of the appearance of the lognormal, our measures indicate that the process is more complicated, as the observation of lognormal distribution can be an artifact of the Digg promotion algorithm. A further consideration is omitted here, but we refer to [35].

VI. Discussion of Friendship Efficiency on Digg

In this section, we discuss the efficiency of friendship relations during content dissemination on Digg. In particular, we evaluate the effectiveness of the friendship network during the spread of information from the following perspectives: 1) Do all friendship pairs have equal performance while disseminating content? 2) Are users of high degree successful in activating their friendship links for content dissemination? 3) Can stories be disseminated multi-hop away from the submitter along social links? Finally, we discuss the decrease of friendship efficiency on Digg.

A. Effectiveness of Digg Friendship Network

The Digg friendship network consists of 6 759 937 friendship links. It turns out that the number of friend pairs that dugg at least one common story is surprisingly low (2% out of the entire friendship links). The remaining 98% of friend pairs never dugg the same story in spite of their friendship relation. Moreover, the 2% active friend pairs do not behave uniformly regarding the number of stories being disseminated between each pair of nodes (denoted by W). As plotted in Fig. 12, most of friend pairs dugg less than ten common stories, while only a few friend pairs reacted on many common content.

Apparently, the 2% friend pairs are more successful than the remaining friends in terms of disseminating content. The next questions arises: *Can we identify the 2% friend pairs in the Digg network by topological properties*?

B. Importance of Friendship Links

Our first approach is to examine the relation between the topological importance of the 2% successful friend pairs (connected by friendship links) and their importance during content dissemination. In graph theory, a good measurement of link importance is the betweenness B_l of a link, which is defined as the number of shortest paths between all possible friend pairs in the network that traverse the link [34, p. 329]. For example, if two clusters of Digg users are connected by one link, this link is considered to be important because stories disseminated from one cluster need to traverse the same link in order to be propagated to the other cluster. In this paper, the importance of a link during content dissemination is evaluated as the amount of common stories dugg by the friend pair connected by that link. A friend pair that dugg many content is considered to be influential during the spread of information.

If the importance of the friendship links during content dissemination is indeed associated with the link betweenness, we would expect a positively correlated relation between the aforementioned two quantities. However, the empirical results from our analysis shows that there is no significant relation between the link betweenness and the importance of the 2% friend pairs while disseminating content ($\rho = 0.00006$). Hence, we conclude that information is not disseminated more effectively along friendship pairs that have higher topological importance.

C. Activation of the Digg Friendship Network

The above section demonstrates that there does not exist a particular type of links (i.e., links with high betweenness) in the Digg friendship network that will effectively disseminate content. However, it may be the case that certain users are very successful in "activating" their fans to forward the information along their social links. Intuitively, users with higher in-degree will be more influential in terms of activating their fans when propagating content. The number of fans that can be activated during content dissemination is referred as to the *active in-degree* of a user.

We will empirically show that such an assumption is not true on Digg. In fact, the activation ratio of a user's direct fans is extremely low. Although there exists a strong correlation ($\rho = 0.76$) between the active in-degree and the total in-degree, a Digg user can only activate 0.7% of his total number of fans to disseminate the content. Hence, there is no significantly influential high-degree users who can activate many fans: a user with 1000 fans is only able to effectively forward the content to 7 fans, on average. We also see that active users who dugg many stories may not be connected to users who are also active in the friendship network (linear correlation coefficient between the number of stories dugg by each friend pair is 0.05).

Another interesting observation from our analysis is that information can indeed traverse multiple hops across the Digg friendship network. However, the spread of information dies out quickly. On average, the content is disseminated no further than 3.9 hops away from the submitter. In fact, nearly 70% of the friend diggers are direct friends of the submitter, while the contribution of multi-hop friend relation during content dissemination decreases exponentially. Hence, it is not surprising that the effective distance of information dissemination over Digg is even smaller than its average shortest path length (5.6 hops).

Both the low activation ratio of a user's direct friends and the non-correlated relationship between the edge betweenness and the strength of a link are examined from a pure topological point of view, and evaluated for the entire Digg friendship network. As we will show in Section VII, there exists another factor that leads to the success of content dissemination along a friendship link: friends should be active on Digg during an appropriate time period.

D. Synchronization of Friends' Activity in Time

As discussed in Section V-BIII, stories must obtain sufficient attention within the first 24 hours after their publication in order to be promoted. Hence, the synchronization of friends' activity in time is critical to the successful promotion of a story along the friendship links. By "activity synchronization", we mean that users are visiting the Digg website and actively digging stories during a proper time slot.

From our dataset, we studied the growing pattern of the entire established friendship links, the active friend pairs (i.e., friends who visited the Digg website and dugg at least one published story), and the active friend pairs who dugg at least one common story over the past four years (from July 2005 to May 2009). As shown in Fig. 13, the number of established friendship links grows almost 100 times from 2005 until 2009 (bin size in months). The number of active friend pairs is significantly smaller (at least two order less) than the actual amount of friend pairs who have registered on Digg.

Since the first 24 hours are critical for stories to get promoted after their publication, we further investigate the probability that an active friends pair will react on the same content on a single day. To do so, we compute the number of friend pairs who are actively visiting Digg.com and digging stories per day over the period since July 2005 to May 2009. On average, there are 196 active friend pairs per day. Besides, we also calculate the number of active friend pairs who have dugg at least one common story per day. Our analysis suggests a strong relation between the amount of active friend pairs and the number of friend pairs digging the same story ($\rho = 0.9$), while the slope of the linear regression line [34, p. 31] is around 0.3, indicating that 30% of the active friend pairs will react on the same content daily.

In Section V-BIII, we have empirically shown that to promote a story via friendship relations, friends have to contribute, on average, 5 diggs per hour; and that the average duration of promoting a story by friends is around 12.6 h; see Fig. 10(a). Thus, for a popular story on Digg, there are, on average, 63 diggs made by active users (connected by friendships) before its promotion. As mentioned earlier, there exist about 196 active friend pairs daily and one third of them digging the same content, which seems to be sufficient for the promotion of new published content. However, considering the huge amount of content (15 000 to 26 000 stories) being published on Digg, the different interests of active friends, and the limited capability of users to digg, it is not surprising that the efficiency of the Digg friendship network for content dissemination decreases. To summarize, synchronization of friends' activity is critical to the effectiveness of the Digg friendship network for the spread of content, especially when disseminating information which becomes obsolete in a short period of time.

The temporary synchronization of digging patterns between random users also results in successful content dissemination by non-friends. Users need to discover, read, and vote for the same story although they are exploring the Digg website via different interfaces. Moreover, stories are shifted to subsequent pages very fast once being published.[18] It is important that users should have found and dugg the same story before it is flooded by the huge amount of stories being submitted to Digg. Therefore, an accurate user behavior model (e.g., users' browsing/digging patterns) on Digg is necessary to predict the group of users who will click on the same story.

[18]Popular stories stay 2 or 3 h on each front page, while upcoming stories are shifted very fast once new stories are published.

VII. Conclusion

This paper presented an empirical analysis of the social media aggregating website, Digg.com. As shown from our study, the novel features on Digg have led to distinct user characteristics and information dissemination pattern. In the following, we highlight our major conclusions.

We crawled the most valid information about the friendship relations, the user activities, and the published content on Digg by employing a simultaneous exploration of the Digg network. Our data collection has shown that the Digg friendship network consists of a large giant component, a number of small connected components, and a significant number of disconnected users. The Digg giant component presents a power law degree distribution, the non-correlated node degree between friends, a remarkably low average shortest path length, and a very low fraction of symmetric links.

We found that the flavor of the Digg community is dominated by a small group of users (2%), as they succeeded in submitting popular stories. The Digg users behave heterogeneously when exploring the Digg content. Users can discover and disseminate content without the engagement of their friendship relations, as the disconnected users on Digg are also actively digging stories. A Digg story needs to attract sufficient attentions after its publication in order to become popular. Once being promoted to the front page, the story quickly obtained a significant amount of votes, as it has been exposed to a vast of audience. The lifetime of the content, on the other hand, is very short. Stories lose their popularity very fast and the number of diggs saturate approximately after 40 h since their promotion.

We have empirically shown that there indeed exists a high overlap between the interests of friends. However, the Digg friendship relations can only successfully disseminate content in half of the cases. In the remaining situations, the Digg content are predominantly promoted by non-friends. We attribute the above observation to the fact that friendship relation is not the only mean for content discovery on Digg and that users are taking advantage of multiple interfaces to explore and further digg content. In addition, we found that only 2% of the entire friend pairs have disseminated the same content. There does not exist a particular group of users who are successful in activating their friends to spread information. Information cannot be propagated further than 3.9 hops away from the submitter in the Digg network. Finally, we argued that the synchronization of friends' digging activities is the key to the success of friendship relations, especially when disseminating content with transient life duration and short attraction period. Information dissemination over Digg friendship links is suppressed, since the digging activities between friends are not always properly synchronized. Hence, designers should take into account of the influence of service interfaces and user behaviors on content dissemination when designing future web applications that aim to share information between end-users.

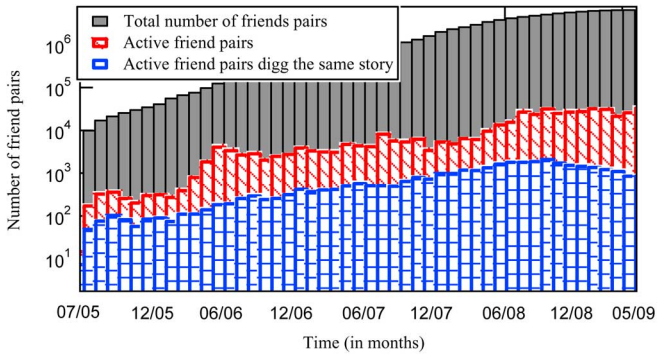

Fig. 13. Number of established friend links, active friend pairs, and active friend pairs digging the same story from July 2005 to May 2009 (bin size in months). The figure is plotted on a log-linear scale for easier reading.

Although the work presented in this paper has been conducted in a thorough and careful way, the applicability of our findings to other types of OSNs (such as Facebook or LinkedIn) needs more investigation. This is because the main function of Digg (i.e., discover and share information) differs from the aforementioned OSNs (i.e., network with friends), as well as the semantics of friendship. Hence, we confine our analysis to OSNs that are similar to Digg. As a future work, comparing the effectiveness of friendship during information dissemination between Digg and other types of OSNs should be performed. In particular, friendship links on Digg present a low level of symmetry, while social connections in other OSNs (e.g., Flickr) are highly symmetric. It would be interesting to examine whether link symmetry plays a role during the spread of content, and whether the semantics of friendship will influence the way that information is disseminated.

ACKNOWLEDGMENT

The authors would like to thank W. Winterbach for proofreading this paper.

REFERENCES

[1] Y. Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong, "Analysis of topological characteristics of huge online social networking services," in *Proc. 16th Int. Conf. World Wide Web*, 2007, p. 844.

[2] L. A. N. Amaral, A. Scala, M. Barthelemy, and H. E. Stanley, "Classes of small-world networks," *Proc. Nat. Acad. Sci. United States of America*, vol. 97, no. 21, p. 11149, 2000.

[3] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan, "Group formation in large social networks: Membership, growth, and evolution," in *Proc. 12th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2006, p. 54.

[4] A. L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[5] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida, "Characterizing user behavior in online social networks," in *Proc. 9th ACM SIGCOMM Conf. Internet Measurement*, 2009, pp. 49–62.

[6] D. M. Boyd and N. B. Ellison, "Social network sites: Definition, history, and scholarship," *J. Comput. Mediated Commun.-Electronic Edition*, vol. 13, no. 1, p. 210, 2007.

[7] M. Cha, H. Kwak, P. Rodriguez, Y. Y. Ahn, and S. Moon, "Analyzing the video popularity characteristics of large-scale user generated content systems," *IEEE/ACM Trans. Netw.*, vol. 17, no. 5, pp. 1357–1370, 2009.

[8] M. Cha, A. Mislove, and K. P. Gummadi, "A measurement-driven analysis of information propagation in the flickr social network," in *Proc. 18th Int. Conf. World Wide Web*, 2009, pp. 721–730.

[9] D. W. Drezner and H. Farrell, "The power and politics of blogs," *Amer. Polit. Sci. Assoc.*, vol. 2, 2004.

[10] H. Ebel, L. I. Mielsch, and S. Bornholdt, "Scale-free topology of e-mail networks," *Phys. Rev. E*, vol. 66, no. 3, p. 35103, 2002.

[11] D. Fono and K. Raynes-Goldie, "Hyperfriends and beyond: Friendship and social norms on LiveJournal," *Internet Res. Annu.*, vol. 4, 2006.

[12] D. Garlaschelli and M. I. Loffredo, "Patterns of link reciprocity in directed networks," *Phys. Rev. Lett.*, vol. 93, no. 26, p. 268701, 2004.

[13] B. A. Huberman and L. A. Adamic, "Growth dynamics of the world wide web," *Nature*, vol. 401, no. 131, 1999.

[14] H. Jeong, S. P. Mason, A. L. Barabási, and Z. N. Oltvai, "Lethality and centrality in protein networks," *Nature*, vol. 411, no. 6833, pp. 41–42, 2001.

[15] J. M. Juran, *Quality Control Handbook*, 3rd ed. New York: McGraw-Hill, 1974.

[16] E. Katz and P. F. Lazarsfeld, *Personal Influence*. New York: Columbia Univ. Bureau of Applied Social Research, 1955.

[17] G. Kossinets, J. Kleinberg, and D. Watts, "The structure of information pathways in a social communication network," in *Proc. 14th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2008, pp. 435–443.

[18] K. Lerman, "Social networks and social information filtering on digg," in *Proc. 1st Int. Conf. Weblogs and Social Media (ICWSM-07)*, 2007.

[19] K. Lerman and R. Ghosh, "Information contagion: An empirical study of spread of news on Digg and Twitter social networks," in *Proc. 4th Int. Conf. Weblogs and Social Media (ICWSM)*, May 2010.

[20] K. Lerman and T. Hogg, "Using a model of social dynamics to predict popularity of news," in *Proc. 19th Int. World Wide Web Conf. (WWW)*, 2010.

[21] J. Leskovec and E. Horvitz, "Planetary-scale views on a large instant-messaging network," in *Proc. 17th Int. Conf. World Wide Web*, 2008, pp. 915–924.

[22] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins, "Geographic routing in social networks," *Proc. Nat. Acad. Sci. United States of America*, vol. 102, no. 33, p. 11623, 2005.

[23] M. O. Lorenz, "Methods of measuring the concentration of wealth," *Publ. Amer. Statist. Assoc.*, pp. 209–219, 1905.

[24] A. Mislove, H. S. Koppula, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Growth of the Flickr social network," in *Proc. 1st Workshop Online Social Networks*, 2008, pp. 25–30.

[25] A. Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *Proc. 7th ACM SIGCOMM Conf. Internet Measurement*, 2007, pp. 42–42.

[26] M. Mitzenmacher, "A brief history of generative models for power law and lognormal distributions," *Internet Math.*, vol. 1, no. 2, pp. 226–251, 2004.

[27] A. L. Montgomery, "Applying quantitative marketing techniques to the internet," *Interfaces*, vol. 31, no. 2, pp. 90–108, 2001.

[28] S. Mossa, M. Barthelemy, H. E. Stanley, and L. A. N. Amaral, "Truncation of power law behavior in scale-free network models due to information filtering," *Phys. Rev. Lett.*, vol. 88, no. 13, p. 138701, 2002.

[29] M. E. J. Newman, "Assortative mixing in networks," *Phys. Rev. Lett.*, vol. 89, no. 20, p. 208701, 2002.

[30] J. Porter, *Designing for the Social Web*. Indianapolis, IN: New Riders Press, 2008.

[31] K. Raynes-Goldie, "Pulling sense out of today's informational chaos: Livejournal as a site of knowledge creation and sharing," *First Monday*, vol. 9, no. 1, 2004.

[32] E. M. Rogers, *Diffusion of Innovations*, 5th ed. New York: Free Press, 2003.

[33] G. Szabó and B. A. Huberman, "Predicting the popularity of online content," *Commun. ACM*, vol. 53, no. 8, pp. 80–88, Aug. 2010.

[34] P. Van Mieghem, *Performance Analysis of Communications Networks and Systems*. Cambridge, U.K.: Cambridge Univ. Press, 2006.

[35] P. Van Mieghem, N. Blenn, and C. Doerr, "Lognormal distribution in the Digg online social network," *Eur. J. Phys. B*, to be published.

[36] P. Van Mieghem, H. Wang, X. Ge, S. Tang, and F. A. Kuipers, "Influence of assortativity and degree-preserving rewiring on the spectra of networks," *Eur. Phys. J. B*, vol. 76, no. 4, pp. 643–652, 2010.

[37] D. M. Wilkinson and B. A. Huberman, "Cooperation and quality in Wikipedia," in *Proc. 2007 Int. Symp. Wikis*, 2007, pp. 164–164.

[38] B. A. Williamson, EMarketer Social Network Marketing: Ad Spending and Usage, 2007. [Online]. Available: http://www.emarketer.com/Report.aspx?code=emarketer-2000478.

[39] F. Wu and B. A. Huberman, "Novelty and collective attention," *Proc. Nat. Acad. Sci.*, vol. 104, no. 45, p. 17599, 2007.

Siyu Tang received the M.Sc. and Ph.D. degrees in electrical engineering at the Delft University of Technology, Delft, The Netherlands, in 2006 and 2010, respectively.

Her research interests include modeling information dissemination in distributed network, performance analysis in peer-to-peer (P2P) streaming systems, and empirical analysis in online social networks. Since 2011, she has been working at Alcatel-Lucent, Antwerp, Belgium, in the IP Division.

Norbert Blenn received the Diplom (equivilant to the M.Sc. degree) in computer science in media from the Technical University of Dresden, Dresden, Germany, in 2007. He is pursuing the Ph.D. degree under supervision of Prof. dr. ir. Piet van Mieghem and Dr. C. Doerr at the Delft University of Technology, Delft, The Netherlands.

In 2010, he joined the Group for Network Architectures and Services, Department of Telecommunications in the faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS) of the Delft University of Technology. His current research interests include content propagation, security, crawling techniques, and the development of online social networks. He is part of the Trans Sector Research Academy for Complex Networks and Services.

Christian Doerr received the joint Ph.D. degree in computer science and cognitive science as well as the M.Sc. degree in computer science from the University of Colorado, Boulder, and a Diplomwirtschaftsinformatik degree from the University of Paderborn, Paderborn, Germany.

He is an Assistant Professor in the Department of Telecommunication at TU Delft, Delft, The Netherlands. His research interests include the empirical analysis of online social networks, the emergent behavior of complex systems, and the resilience of networks.

Piet Van Mieghem received the Master's and Ph.D. degrees in electrical engineering from the K.U. Leuven, Leuven, Belgium, in 1987 and 1991, respectively.

He is a Professor at the Delft University of Technology, Delft, The Netherlands, with a chair in telecommunication networks and Chairman of the section Network Architectures and Services (NAS) since 1998. His main research interests lie in modeling and analysis of complex networks (such as biological, brain, social, infrastructural, etc. networks) and in new Internet-like architectures and algorithms for future communications networks. Before joining Delft, he worked at the Interuniversity Micro Electronic Center (IMEC) from 1987 to 1991. During 1993–1998, he was a member of the Alcatel Corporate Research Center in Antwerp, Belgium. He was a visiting scientist at MIT (Department of Electrical Engineering, 1992-1993) and a visiting professor at UCLA (Department of Electrical Engineering, 2005) and at Cornell University (Center of Applied Mathematics, 2009). He is the author of three books: *Performance Analysis of Communications Networks and Systems* (Cambridge, U.K.: Cambridge Univ. Press, 2006), *Data Communications Networking* (Amsterdam, The Netherlands: Techne, 2006; 2nd ed., 2001), and *Graph Spectra for Complex Networks* (Cambridge, U.K.: Cambridge Univ. Press, 2011).

Prof. Van Mieghem currently serves on the editorial board of the IEEE/ACM TRANSACTIONS ON NETWORKING AND COMPUTER COMMUNICATIONS. He was member of the editorial board of the journal *Computer Networks* from 2005–2006.