Exploring the complex pattern of information spreading in online blog communities

Sen Pei[1,2*], Lev Muchnik[3], Shaoting Tang[1], Zhiming Zheng[1], Hernán A. Makse[2]
[1] Laboratory of Mathematics, Informatics and Behavioral Semantics, and School of Mathematics and Systems Science, Beihang University, Beijing, China
[2] Levich Institute and Physics Department, City College of New York, New York, USA
[3] School of Business Administration, The Hebrew University of Jerusalem, Israel
[*] Corresponding author
E-mail: peisen@buaa.edu.cn (SP)

Abstract

Information spreading in online social communities has attracted tremendous attention due to its utmost practical values in applications. Despite that several individual-level diffusion data have been investigated, we still lack the detailed understanding of the spreading pattern of information. Here, by comparing information flows and social links in a blog community, we find that the diffusion processes are induced by three different spreading mechanisms: social spreading, self-promotion and broadcast. Although numerous previous studies have employed epidemic spreading models to simulate information diffusion, we observe that such models fail to reproduce the realistic diffusion pattern. In respect to users behaviors, strikingly, we find that most users would stick to one specific diffusion mechanism. Moreover, our observations indicate that the social spreading is not only crucial for the structure of diffusion trees, but also capable of inducing more subsequent individuals to acquire the information. Our findings suggest new directions for modeling of information diffusion in social systems and could inform design of efficient propagation strategies based on users behaviors.

Introduction

Information spreading is a key social phenomenon, which due to the prevalence of communication and information technologies in the recent years is becoming increasingly dynamic and powerful in shaping such processes as the adoption of innovations [1], the propagation of news and information [2–4], the success of viral marketing [5–8] as well as the spread of behaviors and social norms in general [9, 10]. With the mainstream adoption of Internet and World Wide Web, several types of user-based information dissemination platforms, such as the blogs sharing communities, online social networks, and microblog services, have become prevalent. In particular, the blogspace has attracted millions of users to discuss or share topics ranging from public concerned affairs to users' personal lives, thus greatly facilitating people to create, capture and disseminate information [11–14]. Indeed, people's life and social relations have been influenced to a great extent by the information in the blogspaces. As a consequence, understanding the mechanism by which a piece of information diffuses through population is crucial for not only designing efficient promotion strategies but also blocking the pervasion of malicious information. Because of its practical values in real-world applications, information spreading in online communities has attracted much attention across disciplines. In particular, many studies have been addressing to model the information spreading process [15–17].

Much previous research investigating the information spreading among individuals has been motivated by analogy with the spread of a contagious disease [18–22]. In these disease-propagation models, a piece of information is assumed to diffuse along social links in the underlying social network from person to person complying with specific rules and may eventually reach a large fraction of the population [23]. With these assumptions, it has been shown that the social network structure can significantly affect the outcome of a spreading process [19, 20, 24]. Although such epidemic-motivated approaches have led to

many profound results, these models are typically based on rather simplified assumptions that may not be representative for information spreading in real circumstances. Recent observational studies tracking actual diffusion processes demonstrate that information and diseases spread differently: apart from the network structure, the contagion of information is also affected by other factors, including the human behavior [25–27], homophily [28, 29], social reinforcement [10], etc. With the existence of such factors, the diffusion paths of information in real world are found to be dramatically different from those derived from the classical epidemic models [30]. In addition, recent empirical studies on information flows in blogspaces have presented evidences that information does not necessarily diffuse via social links [31–33], which opposes to the prevailing assumption that information spreading is confined within the underlying social networks. All these observations of real-world diffusion instances indicate that we still lack the systematic understanding of how information diffuses in online social communities.

In this paper, we perform a detailed analysis of information spreading in an online blog community - LiveJournal. Earlier studies of the individual-level diffusion patterns in blogspaces [11–14] focused on the topological feature of diffusion paths. Here, by analyzing information diffusion trees along with the social network structure, we seek to unveil the diffusion pattern of information spreading in blog communities based on users' behaviors of disseminating information. We find that the spreading instances can be classified into three types: *social spreading*, which occurs following social links, *self-promotion*, meaning references of earlier posts by the same author, and *broadcast*, representing the remaining cases. A typical spread of information in social network can therefore be seen as a combination of these three processes. We explore the topological structure associated with each of the three categories and find that the vast majority of diffusion cases are confined to just first few steps. Through simulations of the susceptible-infected-recovered (SIR) model with real-world infection rate, we find that the SIR model fails to reproduce the realistic social spreading pattern. This distinguishes information diffusion in social networks from the epidemic spreading. In respect of the adoption of specific behavioral pattern by users, it is striking that nearly all the users would stick to an inherent diffusion type. Such persistence in behavior of the individual users can be interpreted as manifestation of their preferences and the role they pursuit in the social network. We proceed by further investigating the dynamical coupling of distinct diffusion mechanisms with a particular focus on social spreading. Although, social spreading is observed less frequently than the other mechanisms, it tends to be associated with larger cascades and can induce more subsequent disseminations. This finding implies that, with the propelling of social spreading, information could reach a much larger fraction of the population, which highlights the essential role of social spreading in information diffusion. From the perspective of human behaviors, our findings reveal a set of complex behavioral patterns exploiting different information spread mechanisms and suggest new directions for modeling information spreading in online social networks.

Materials and Methods

In order to explore the information diffusion in LiveJournal.com, we have obtained approximately 56 million posts published during a period of 21 months. The social network constructed by friend lists consists of 9,573,127 users and 188,240,039 social links. Aimed to track the information flow from one user to the other, we analyze the hyperlinks contained in the posts. To be precise, for each post containing hyperlinks to other posts in LJ, we record the following information: ID of the post P_d, ID of the post's author I_d, publication time of the post T_d, ID of the referred post P_s, ID of the referred post's author I_s and the time of the referred post's publication T_s. In total, we have identified 3,462,504 posts referencing other posts in our dataset. Using this data, we are able to track the information flow as follows: user I_d has published a post P_d at time T_d, which contains the hyperlink to post P_s published by user I_s at time T_s. Therefore, the information diffuses from the source user I_s to the destination user I_d explicitly.

The posts data is composed of public posts transmitted by the company hosting the platform via the live stream (http://www.livejournal.com/stats/latest-rss.bml). We do realize that this data, even if

classified as public and freely accessible online, may potentially contain private information. Our study is therefore limited to the exploration of the hyperlinks between the posts, rather than analysis of their textual content. Furthermore, we anonymize the dataset by replacing user names with numerical IDs. A complementary data, the social network structure, was obtained by crawling the LiveJournal web site via the publicly available API, FOAF (http://www.livejournal.com/bots/) designed specifically for that purpose. We would like to note that all social network connections are public. Still, we anonymize the network substituting LiveJournal user names with numerical IDs.

Results

In pursuing to unveil the patterns of information spreading in online social communities, we have collected the complete social network structure as well as the available blog posts of a large-scale online social community - LiveJournal.com (LJ). The details of collected data are explained in Materials and Methods. The LJ platform facilitates maintenance of friend lists that help users track information updated by their peers, thus facilitating information spreading among them. The resulting social network, which consists of nearly 9.6 million users, records the complete information of friend relations in the LJ community. This network composed of social links has been explored in several previous studies and is believed to reliably represent actual social relations in the community [34, 35]. In this work we leverage the LJ lists of followers to construct an undirected graph and use it to track and study propagation of information in that social network. We identify and track the propagating pieces of information by analyzing the content of posts. In LJ community, posts typically refer (by heperlinking) the source of the information which they retransmit, discuss or refer to. By extracting these hyperlinks and matching them to earlier posts, we can directly track the information flow from one user to the other (details in Materials and Methods). This inference technique confines the information flow to the LJ community and excludes the diffusion content coming from external sources, such as news channels and newspapers.

By analyzing the relationship between the source user I_s and destination user I_d in each spreading instance, the spreading pattern can be categorized into three groups. We attribute the content spreading between the users explicitly connected in social network (I_s and I_d are linked) to social spreading by reasoning that such propagation is induced by the underlying social network. We classify the posts citing same user's earlier posts ($I_s = I_d$) as self-promotion. The third, broadcast news propagation pattern, encompasses the posts citing earlier publications of remote (in terms of the underlying social network) users. The broadcast pattern typically relies on search, promotion and the LJ recommender engine to discover the content from remote users. In the collected data, these three types of information diffusion, i.e., the social spreading, self-promotion, and broadcast, make up 26.8%, 31.14%, and 42.06% of the total propagation respectively. This suggests that, contrary to the common belief that information flow is confined to the social network, the real information propagation in online social community exhibits a complex pattern in which the mentioned three diffusion patterns coexist and entwine with each other.

Compared with other online social systems, such as Twitter where the social influence is more than 70% [36], the composition of diffusion links of social spreading is only 26.8%. The difference may come from the features of LJ system. LiveJournal is a social networking service where users can keep a blog, journal or diary. It not only has the typical social function of "friend list", but also maintains large numbers of communities, which are collective blogs in which different users can post messages. Users who are interested in a particular subject can find or create a community for this subject. Notice that users in the same community are not necessarily connected by friend links. Therefore, even though users repost information within one community, these diffusion links will still be classified as broadcast. Our data indicates that over 40% diffusion processes are attributed to such broadcast. And the proportion of social spreading is therefore squeezed to 26.8%.

The named diffusion patterns represent distinct propagation mechanisms. Even when social spreading is considered, the information diffusion processes differ fundamentally from the classical viral contagion.

The information recipient exposed to the content is the one who decides whether to adopt the behavior and share the information among her own local social network. The authors cannot control such kind of passive diffusion. As has been pointed recently [30], the behavior adoption is not completely consistent with the classical epidemic spreading processes, in which the viruses are capable of self-replicating and propagating themselves proactively. Unlike social spreading, the self-promotion mechanism relies on the dedication of the authors to repeatedly promote their own content, increasing exposure and the probability of consequent sharing (i.e. see complex contagion [37]). In contrast to social spreading and self-promotion, broadcast is more likely to be caused by the mechanism of marketing or mass media. These arguments show that, although the viral disease-like diffusion has been intensively explored in previous literature [38], the dynamics of such coupled information spreading processes is still remained to be explored.

Construction of diffusion trees

Aiming to acquire the individual-level details of information spreading, we have reconstructed the exact diffusion tree for each source information. More specifically, we start by sorting the posts by their publication time T_s so that the cited posts are guaranteed to appear chronologically in the record list. Then we identify the source information which has only been referred by other posts but contains no hyperlinks to other posts. For each of these source information, we create a search queue to find the subsequent diffusion in a Breadth-First-Search (BFS) fashion: push the IDs of the posts citing the source information into the search queue; then pop out the front element of the queue, and push into the queue the IDs of the posts that cited the popped-out post but are still not visited in previous steps; this process is applied recursively until the search queue becomes empty. We define the diffusion depth to be the number of layers the information spreads out from the origin. And the size of a diffusion tree is defined as the number of posts found in our search. In total, we have reconstructed 880,195 information diffusion trees in LJ community. Figure 1a presents one real diffusion tree of depth 8 and containing 227 nodes. Figure 1b shows the distributions of the size and depth of the observed diffusion trees. These two quantities exhibit power-law distributions with exponents γ, characterized by $\gamma_s = 1.86 \pm 0.05$ and $\gamma_d = 2.97 \pm 0.29$ correspondingly (the values of γ are obtained using the maximum likelihood method [39]).

The power-law distribution of the diffusion tree size and depth implies that the majority (63.0%) of the posts get to be cited only once. In the posts that acquire numerous citations, diffusion process is fueled by three distinct mechanisms, as can be seen from Fig. 1a. In particular, the links representing social spreading, self-promotion, and broadcast are marked with different colors. It is the complex dynamical interaction between these three kinds of diffusion mechanisms that results in, perhaps infrequent, but significant cases of information propagation in social network. We continue by detailed exploration of each of the three diffusion patterns and the interaction between them.

Social spreading

In this section, we examine how information propagates via the underlying social network. Much of the current research perceives viral spread of information in full analogy with the spread of contagious diseases [5,6]. This has led to the wide adoption and extensive use of the classical disease models like susceptible-infectious-recovered (SIR) and susceptible-infectious-susceptible (SIS) in information propagation studies [18–21,24].

As we focus on *social contagion* in detail, we eliminate the other types of information propagation links and obtain 363,115 distinct diffusion trees entirely composed of social spreading links. The distribution of tree size still follows a power-law shape, although with a larger exponent of $\gamma = 2.16 \pm 0.11$ (Fig. 2a). Even though, exceptionally deep social diffusion trees do indeed occur, they are too rare to play significant role in information spread processes. We find that over 85% of social spreading is attributed to cascades that do not exceed the depth of 3 (Fig. 1b). Therefore, in online blog communities, the majority

of social spreading occurs via small and shallow information cascades. Our observation is in accordance with previous findings in other online communities [30]. Recall that for epidemic spreading, infections over multiple generations are responsible for most of the contagion. This differs fundamentally from our observation in information spreading. Such difference motivates us to further explore the topological structure of diffusion trees. To delve into this issue, we examine the branching number (number of children) of each post in diffusion trees. More specifically, we identify posts' depth in the diffusion and then display the average branching number for each depth in Fig. 2d. The average branching number of nodes in the first few generations varies above 1, while for posts located at the depth of more than 20, the branching number is almost 1. As a consequence, for extremely deep diffusions, most of the posts are shared within the first few generations.

Now we turn to explore the relationship between the spreading dynamics and underlying social networks. First, we map the diffusion trees in which nodes represent posts onto the underlying social network in which nodes are individuals. For simplicity, we will refer the spreading among users following social links as viral spreading in our discussion. To avoid unnecessary complexity we represent repeated referrals to the same post by the same user with a single link corresponding to the earliest citation. The obtained viral spreading is therefore represented by a directed graph rather than a tree. Furthermore, numerous appearances of the same user in the diffusion tree translate to a single node in the viral spreading graph, resulting in substantially smaller constructs. Figure 2c shows the relation between the viral spreading size and diffusion trees' size. The diminishing ratio, which is defined as the ratio between the size of viral spreading and corresponding diffusion trees, decreases significantly as the tree size grows. This means, for larger diffusion trees, there will be more users repost same information repeatedly. Consequently, larger diffusion trees may not necessarily reach larger population. The distribution of viral spreading size is still a power-law, with an exponent $\gamma = 2.26 \pm 0.12$, while the distribution of viral spreading depth remains almost the same as that of diffusion trees (see Fig. 2a).

The representation of viral spreading graphs helps confirming that most of the diffusion instances are indeed confined to very few steps. In agreement with our earlier observation in diffusion trees, we find that over 95% of spreading instances occur in one to three layers, as shown in Fig. 2b. If we classify the nodes according to their depth in the viral spreading, we find that for the nodes with depth less than 10, the average branching number remains above 1 (Fig. 2d). On the contrary, all the nodes deeper than 10 generations have the branching number of 1. This phenomenon indicates that most of the viral spreading occurs in the first few layers.

To better explore the difference between the pattern of information and epidemics spreading, we perform extensive SIR simulations with real-world infection rate. In SIR model, infected nodes infect their susceptible neighbors with probability β and then enter the recovered state with probability λ, where they become immunized and cannot be infected again [21, 24]. In our study, we set $\lambda = 1$. For users involved in viral spreading, we define their infection rate β as the fraction of neighbors who repost their information. If a user participates in more than one spreading instance, we just adopt the average value of individual infection rate. Unlike the common assumption that individuals have same infection rate in previous studies, the distribution of realistic infection rate is quite heterogeneous (Fig. 3a). Starting from the same spreading sources as empirical viral spreading, we conduct 100 SIR realizations for each source with realistic infection rate. We take the average size and depth of these realizations as the outcomes of SIR simulations. In Fig. 3b, we display the ratio of size probability density between SIR simulations and real viral spreading P_{SIR}/P_{real}. SIR model overestimates the number of moderate diffusion, but underestimates the number of size 1 and extremely large diffusion. For spreading depth, SIR model also bias to diffusion with medium depth (see inset of Fig. 3b). Since SIR modeling and real-world viral spreading have same information sources and infection rate, the discrepancy should attribute to the underlying spreading mechanism.

We further examine the structural difference between real viral spreading and simulated epidemic diffusion. In Fig. 3c, we show the proportion of diffusion trees with a certain depth. The ratio between

real cases and SIR model is displayed in the inset. Compared with SIR model results, about 90% realistic viral spreading only last for one generation. On the contrary, nearly 50% SIR diffusion trees have multiple-step cascades. For diffusion trees with a certain depth, we show the proportion of spreading links in these trees in Fig. 3d and the ratio between real viral spreading and SIR simulations in the inset. While multiple-step cascades are responsible for the majority (60%) of spreading links in SIR model, over 80% spreading links are limited to one-step diffusion in real viral spreading. On the whole, SIR model overestimates the significance of multiple-step cascades and fails to reproduce the spreading pattern in real scenario. Therefore, our observation questions the previous adoption of epidemic-driven models in the research of information diffusion.

Since the social spreading pattern is coupled with other types of spreading patterns in the process of information diffusion, and the links in social spreading pattern may be affected by other spreading types even though the other links are deleted, the difference between social spreading pattern may come from the interaction among the spreading pattern. To make clear of this point, we eliminate the effect of the interaction among the spreading patterns and compare the real information diffusion with SIR model. We extract 231,333 diffusion trees entirely composed of social spreading pattern, which have no interaction with other diffusion types. The analysis results are shown in Fig. 3e and Fig. 3f. Clearly, there exists fundamental discrepancy between real diffusion and SIR model. Therefore, the difference should stem from the spreading mechanism, rather than the interaction among spreading patterns.

In the above experiment, we perform SIR model with individuals' infection rate defined by the fraction of neighbors who repost one's information. However, in general, each user should exert different influence on distinct neighbors. Considering this fact, we also conduct SIR simulation with infection rate inferred for each link. For a given directed social link, the infection rate is calculated as follows: we divide the number of diffusion instances from the source node to the end node by the total number of posts published by the source node in all social spreading trees. The distribution of links' infection rate is displayed in Fig. 4a, which is also heterogeneous. We perform same analysis for SIR model with links' infection rate and present the results in Fig. 4b-f. For both viral spreading with and without interaction with other diffusion types, SIR model with links' infection rate cannot reproduce the realistic viral spreading pattern - it also overestimates the importance of multiple-step cascades.

Self-promotion

The diffusion type of self-promotion makes up a large fraction of the information diffusion. Self-promotion enhances the exposure of information, thus increasing the potential of being cited by other posts. Among the 315,937 individuals participating in information spreading, only 53,835 users have the behavior of self-promotion. We reconstruct the diffusion trees of self-promotion as before and obtain 254,861 diffusion trees. The distribution of the size and depth of these trees are displayed in Fig. 5a. Both of them follow pow-law distributions approximately. However, compared with the diffusion trees of social spreading, the maximum size and depth are much smaller. For instance, the maximum depth of social spreading can be up to several hundreds, while the maximum depth is only 85 in case of self-promotion. Among all the self-promotion instances, 92.13% are contributed by the diffusion trees that last less than 5 generations (see Fig. 5b). Different from social spreading in which over 80% spreading links locate in one-step diffusion, self-promotion trees with depth 2 contribute above 50% of all spreading links. Intuitively, this phenomenon can be explained by the behavioral preference of each diffusion type. For social spreading, people usually place more trust in their direct neighbors. Therefore, most social spreading should concentrate in the first layer. On the contrary, if users decide to self-promote their own posts, they usually do not stop at the original ones. They would continue self-promote their subsequent posts, attempting to make more exposures. That may be the reason why most self-promotion resides in multi-step diffusion trees.

When we map the self-promotion trees to the diffusion among users, the process is simple since each self-promotion tree corresponds to a unique user. Figure 5c shows that the total number of self-promotion

for each user follows a power-law distribution with an exponent $\gamma = 1.62 \pm 0.08$. This means while most of the users perform self-promotion for very few times, there exist a small number of users self-promote their posts frequently. If we plot the relationship between posts' branching number and their depth in the self-promotion diffusion trees, we find that while posts in the first few layers can have an either extremely large or small branching number, most of the posts located deep in the trees have branching number of 1. Therefore in case of self-promotion, users are more motivated to repost their earlier posts, which appears in the first few generations in the self-promotion diffusion trees.

Broadcast

Similar with the analysis of social spreading and self-promotion, we reconstruct 441,027 diffusion trees of broadcast. As shown in Fig. 6a, the distribution of the size and depth of these trees are power-law, with exponents $\gamma = 1.98 \pm 0.08$ and $\gamma = 2.86 \pm 0.33$ separately. Although the distribution is similar with that of social spreading and self-promotion, there exists a remarkable difference in the constitution of broadcast diffusion trees. From Fig. 6b, we can conclude that a considerable fraction of broadcast occurs in diffusion trees last for many generations, which opposes to the cases of social spreading and self-promotion. Considering the power-law distribution of tree depth, our observation indicates that even though the deep diffusion trees are rare, their scales are extremely large so that the broadcast links in these trees can occupy a significant fraction of the total broadcast instances. This structural difference may highlight the crucial role of broadcast playing in the information diffusion in LJ community. In Fig. 6d, the relationship between the branching number and nodes' depth also holds for broadcast, i.e. the branching number for nodes in less than 20 layers varies in a wide range and has mean values larger than 1, whereas nodes deeper than 20 generations have only one subsequent broadcast.

Since one user can perform broadcast for many times, the population participating in a broadcast diffusion tree should be smaller than the tree size. We denote the spreading processes among users in broadcast as broadcast spreading. In Fig. 6c, we display the relationship between the size of broadcast spreading and corresponding diffusion trees. Compared with the case of social spreading, the diminish ratio is higher for deep diffusion trees. This means in broadcast the diffusion trees are capable of reaching larger population under the same circumstances. The size and depth of broadcast spreading are also power-law distributed, with exponents $\gamma = 2.04 \pm 0.08$ and $\gamma = 2.90 \pm 0.35$ respectively(see Fig. 6a). In broadcast spreading, 94.91% of the involved users repost information for the first time in less than 4 steps, as displayed in Fig. 6b. Recall that for broadcast diffusion trees, deep ones are responsible for a large fraction of broadcast, our finding implies that after users have reposted the information for the first time, they still keep citing the information in other posts so that the diffusion trees can grow deeper. Same as previous analysis, we plot the branching number of nodes versus their depth in the broadcast spreading in Fig. 6d. Deep nodes only lead to one subsequent propagation. Most of the users adopt the information in just a few generations.

Human activity

One critical factor affecting the spreading outcome is the human activity, including activity frequency [27] and response time [25]. To be precise, the activity frequency is defined as the number of each user's participation in a specific diffusion type, such as social spreading, self-promotion and broadcast. For example, if a user is observed to have cited n posts of his/her neighbors, his/her activity frequency of social spreading is defined as n. In addition, the response time is defined as the time it takes for an individual to repost the information. For instance, if at time t_1 a user cited a post which was published at time t_0, the response time of this diffusion instance is $t_1 - t_0$. In Fig. 7a, we display the distributions of users' activity number of social spreading, self-promotion, broadcast as well as the overall diffusion. They are extremely similar except in the range of tails. The maximum number of social spreading is

much smaller than that of self-promotion and broadcast, which may be due to the existence of robot users who automatically repost their own publications or gather other users' popular posts.

One may wonder if users tend to conduct all types of diffusion or just incline to adopt only one type. In order to answer this question, for each user, we calculate the fraction of each diffusion mechanism. For instance, if a user performs a_1 social spreading, a_2 self-promotion, a_3 broadcast and a total of $A = a_1 + a_2 + a_3$ diffusion, then the fraction for these three types are a_1/A, a_2/A, and a_3/A. We show the distribution of the obtained fraction for each type in Fig. 7b. A shocking phenomenon is that for all the cases, the fraction concentrates near the values of 0 and 1, while the intermediate values are relatively rare. This implies the spreading behavior of users is highly personalized: most of the users choose to adopt only one specific spreading behavior. Therefore, our classification of spreading types is related to users' inherent behaviors. Apart from users preferring to repost their peers' posts, there exist considerable numbers of users who only adopt self-promotion and broadcast. The function of these users in information spreading will be discussed in the next section.

Another important factor of human activity is the response time τ [40]. In Fig. 7c, we plot the distribution of response time for each diffusion type, adopting the time unit of second in the main panel and day in the inset. The response time ranges over several magnitude, which exhibits extremely heterogeneity. Vast majority of spreading takes place within the time interval between 0.1 to 10 days. Therefore the characteristic time of information spreading in LJ is about one day, which is far smaller than the time scale of the formation of network structure. If we change the time unit from second to day, all three types of diffusion follow a similar power-law distribution. Previous research has shown that the power-law distribution of human activity can have a significant impact on information spreading [25].

Dynamical coupling of different spreading mechanisms

Since the diffusion trees of posts are induced by the combination of distinct mechanisms, it is desirable to investigate how these spreading types couple with each other. First we should check the fraction of each type in the diffusion trees. On the whole, the diffusion trees are composed of 20% social spreading, 18.3% self-promotion, and 61.7% broadcast. Note that this composition is different from that of diffusion links, which is 26.8%, 31.14%, and 42.06%. The reason for this discrepancy lies in that posts may contain several hyperlinks to other posts. As a consequence, a single spreading instance could participate in more than one diffusion trees. On average, broadcast posts contain more referred hyperlinks, which raises the proportion of broadcast. In order to check the composition of diffusion trees with different depths, we present the proportion of the social spreading, self-promotion, and broadcast links in diffusion trees deeper than a given threshold in Fig. 8a. We first select diffusion trees whose depth exceeds a given value, and then calculate the fraction of each mechanism in these trees. As the lower bound of depth increases, the fraction becomes steady. This indicates that the composition of diffusion trees is similar for each depth.

To quantify the importance of spreading links in diffusion trees, we define the route number $r(i, j)$ for link (i, j) as the number of shortest paths from the information source to the other nodes passing through this link:

$$r(i,j) = \sum_{t \in V \setminus \{s\}} \sigma_{st}(i,j), \tag{1}$$

where V is the set of nodes in a diffusion tree and $\sigma_{st}(i, j)$ is the number of shortest paths between the source node s and another node t which pass through link (i, j). Apparently, $r(i, j)$ stands for the number of subsequent diffusion induced by link (i, j). The distributions of route number for all the types are displayed in Fig. 8b, which are all highly skewed. In Fig. 8c, with the increase of the lower bound of selected trees' depth, the average route number $\langle r \rangle$ of each mechanism first grows rapidly and then becomes steady, except that $\langle r \rangle$ of social spreading still grows at a slower pace. Although self-promotion

is fewer than social spreading and broadcast, the average route number is larger, which reflects the significant role of self-promotion in the formation of diffusion trees.

Aimed to explore the formation of diffusion trees, we remove the leaves of diffusion trees and check the composition of the obtained *diffusion skeletons*. We focus on this structure since the diffusion trees are expanded based on the diffusion skeletons. In fact, 85.24% nodes located in the leaves are induced by the other 14.76% nodes in diffusion skeletons. Therefore these skeletons determine the top-level structure of diffusion trees. In the inset of Fig. 8d, we can see self-promotion is responsible for up to 60% links in diffusion skeletons. This implies that large numbers of social spreading and broadcast links are in the leaves of diffusion trees, thus pulling down their average route numbers. This would lead to the higher average route number of self-promotion in Fig. 8c . If we check the average route number in diffusion skeletons, surprisingly, social spreading has a much higher value (see Fig. 8d). This indicates that social spreading is crucial in the formation of diffusion skeletons. Although the number of social spreading links is smaller, they incline to locate near the information source, which could induce more subsequent links in skeletons.

As we have pointed out, larger diffusion trees may not necessarily lead to more influenced population. Considering in practice people care more about the number of influenced individuals, we need to explore the dissemination among users. In the following analysis, we map the diffusion trees to spreading among users as we did for social spreading and broadcast. The obtained user spreading only contains links of social spreading and broadcast, since self-promotion links are not capable of producing new adopters. As can be seen in Fig. 8e, broadcast links still hold the majority of user spreading, regardless of our selection of the lower bound for user spreading's depth. The average route number of social spreading is higher than that of broadcast in Fig. 8f. This indicates that, despite that broadcast links' number is higher, social spreading could on average lead to more adopters. This observation highlights the function of social spreading in expanding influenced population.

One may wonder if the route number is related to the spreading speed. Therefore we display the average route number versus response time for three spreading patterns in Fig. 9. For social spreading, it is striking that fast spreading instances with small response time have extremely large route numbers. This implies, on average, the faster one reposts his/her neighbors' information, the more subsequent diffusion he/she will induce. While for self-promotion, it is slow spreading that has larger route number. For broadcast, the average route number has no clear relationship with spreading speed, except a downward trend with growing response time. How these differences arise is an interesting topic to be further explored.

Discussion

Understanding human behaviors associated with dissemination of information and the resulting information propagation patterns in online social communities is crucial for a wide range of applications. Despite the vast and growing literature on information spreading, the relationship between diffusion and users' dissemination patterns has not been explored. Here, we perform a detailed analysis on the diffusion data and social network structure of an online blog community. To our surprise, we find that most users exhibit persistent behavior following one of the three patterns - social spreading, self-promotion and broadcast. We study diffusion trees of each type and show that majority of cases of information propagation are limited to the first few generations. In particular, we compare the spreading pattern of real information diffusion and SIR model through simulations with realistic infection rate. The discrepancy in spreading pattern indicates that widely used epidemic models are incapable of reproducing realistic information spreading, which necessitates more accurate information spread models. Moreover, the claim about the prominence of social spreading in information dissemination is also supported by the fact that social spreading can lead to more subsequent individuals acquiring the information.

The suggested classification of the information diffusion patterns can in future research be applied to

other online social networks, including Facebook and Twitter. Further research is necessary to understand how each of the mechanisms associates with user traits, content and the outcomes of the information diffusion. We believe, that our classification can be employed in new generation of information spread models and for practical use, such as marketing to control information flow in social networks.

Acknowledgments

We are grateful for the suggestions from anonymous reviewers. SP, ST and ZZ are supported by Major Program of National Natural Science Foundation of China (No. 11290141), NSFC (No. 11201018), International Cooperation Project No. 2010DFR00700, Fundamental Research of Civil Aircraft No. MJ-F-2012-04. SP also acknowledges support from Innovation Foundation of BUAA for PhD Graduates.

References

1. Rogers EM. Diffusion of Innovations. 4th ed. New York: Free Press; 1995.

2. Watts DJ, Dodds PS. Influentials, networks, and public opinion formation. J Cons Res. 2007; 34: 441-458.

3. Liben-Nowell D, Kleinberg J. Tracing information flow on a global scale using Internet chain-letter data. PNAS. 2008; 105: 4633-4638.

4. Muchnik L, Aral S, Taylor SJ. Social Influence Bias: A Randomized Experiment. Science. 2013; 341: 647-651.

5. Watts DJ, Peretti J, Frumin M. Viral marketing for the real world. Harvard Business Rev. 2007; May: 22-23.

6. Leskovec J, Adamic LA, Huberman BA. The dynamics of viral marketing. ACM Transactions on the Web (TWEB). 2007; 1: 5.

7. Van den Bulte C, Joshi YV. New product diffusion with influentials and imitators. Market Sci. 2007; 26: 400-421.

8. Aral S, Walker D. Tie Strength, Embeddedness, and Social Influence: A Large-Scale Networked Experiment. Manage Sci. 2014; 60: 1352-1370.

9. Christakis NA, Fowler JH. Connected: The surprising power of our social networks and how they shape our lives. Hachette Digital, Inc. 2009.

10. Centola D.The spread of behavior in an online social network experiment. Science. 2010; 329: 1194-1197.

11. Gruhl D, Liben-Nowell D, Guha RV, Tomkins A. Information diffusion through blogspace. Proc 13th Intl WWW Conf. 2004; pp. 491-501.

12. Java A, Kolari P, Finin T, Oates T. Modeling the spread of influence on the blogosphere. Proc 15th Intl WWW Conf. 2006; pp. 22-26.

13. Gallos LK, Rybski D, Liljeros F, Havlin S, Makse HA. How people interact in evolving online affiliation networks. Phys Rev X. 2012; 2: 031014.

14. Guille A, Hacid H, Favre C, Zighed DA. Information diffusion in online social networks: A survey. ACM SIGMOD Record. 2013; 42: 17-28.

15. Lü L, Chen DB, Zhou T. The small world yields the most effective information spreading. New Journal of Physics. 2011; 13: 123005.

16. Liu C, Zhang ZK. Information spreading on dynamic social networks. Commun. Nonlinear Sci. Numer. Simulat. 2014; 19: 896-904.

17. Zhu YX, Zhang XG, Sun GQ, Tang M, Zhou T, Zhang ZK. Influence of Reciprocal Links in Social Networks. PLoS ONE. 2014; 9: e103007.

18. Hethcote HW. The mathematics of infectious diseases. SIAM Rev. 2000; 42: 599-653.

19. Barrat A, Barthelemy M, Vespignani A. Dynamical processes on complex networks. Cambridge: Cambridge University Press; 2008.

20. Pei S, Makse HA. Spreading dynamics in complex networks. J Stat Mech. 2013; 12: P12002.

21. Kitsak M, Gallos LK, Havlin S, Liljeros F, Muchnik L, Stanley HE, et al. Identification of influential spreaders in complex networks. Nat Phys. 2010; 6: 888-893.

22. Yan S, Tang S, Pei S, Jiang S, Zhang X, Ding W, et al. The spreading of opposite opinions on online social networks with authoritative nodes. Physica A. 2013; 392: 3846-3855.

23. Kleinberg J. Cascading behavior in networks: Algorithmic and economic issues. In: Algorithmic Game Theory. Cambridge: Cambridge University Press; 2007. pp. 613-632.

24. Pastor-Satorras R, Vespignani A. Epidemic spreading in scale-free networks. Phys Rev Lett. 2001; 86: 3200-3203.

25. Iribarren JL, Moro E. Impact of human activity patterns on the dynamics of information diffusion. Phys Rev Lett. 2009; 103: 038702.

26. Funk S, Salath M, Jansen VA. Modelling the influence of human behaviour on the spread of infectious diseases: a review. J R Soc Interface. 2010; 7: 1247-1256.

27. Muchnik L, Pei S, Parra LC, Reis SD, Andrade Jr JS, Havlin S, et al. Origins of power-law degree distribution in the heterogeneity of human activity in social networks. Sci Rep. 2013; 3: 1783.

28. Centola D. An experimental study of homophily in the adoption of health behavior. Science. 2011; 334: 1269-1272.

29. Aral S, Muchnik L, Sundararajan A. Engineering Social Contagions: Optimal Network Seeding in the Presence of Homophily. Netw Sci. 2013; 1: 125-153.

30. Goel S, Watts DJ, Goldstein DG. The structure of online diffusion networks. Proc 13th ACM Conf on Electronic Commerce. 2012; pp. 623-638.

31. Grabowicz PA, Ramasco JJ, Moro E, Pujol JM, Eguiluz VM. Social features of online networks: The strength of intermediary ties in online social media. PLoS ONE. 2012; 7: e29358.

32. Pei S, Muchnik L, Andrade Jr JS, Zheng Z, Makse HA. Searching for superspreaders of information in real-world social media. Sci Rep. 2014; 4: 5547.

33. Li W, Tang S, Pei S, Yan S, Jiang S, Teng X, et al. The rumor diffusion process with emerging independent spreaders in complex networks. Physica A. 2014; 397: 121-128.

34. Backstrom L, Huttenlocher D, Kleinberg J, Lan X. Group formation in large social networks: membership, growth, and evolution. Proc 12th ACM SIGKDD Intl Conf on Knowledge Discovery and Data Mining. 2006; pp. 44-54.

35. Liben-Nowell D, Novak J, Kumar R, Raghavan P, Tomkins A. Geographic routing in social networks. PNAS. 2005; 102: 11623-11628.

36. Myers SA, Zhu C, Leskovec J. Information diffusion and external influence in networks. Proc 18th ACM SIGKDD Intl Conf on Knowledge Discovery and Data Mining. 2012; pp. 33-41.

37. Centola D, Macy M. Complex contagions and the weakness of long ties. Am J Sociol. 2007; 113: 702-734.

38. Pastor-Satorras R, Castellano C, Van Mieghem P, Vespignani A. Epidemic processes in complex networks; 2014. Preprint. Available: arXiv:1408.2701.

39. Clauset A, Shaliz CR, Newman MEJ. Power-law distribution in empirical data. SIAM Rev. 2009; 51: 661.

40. Barabási AL. The origin of bursts and heavy tails in human dynamics. Nature. 2005; 435: 207-211.

Figure Legends

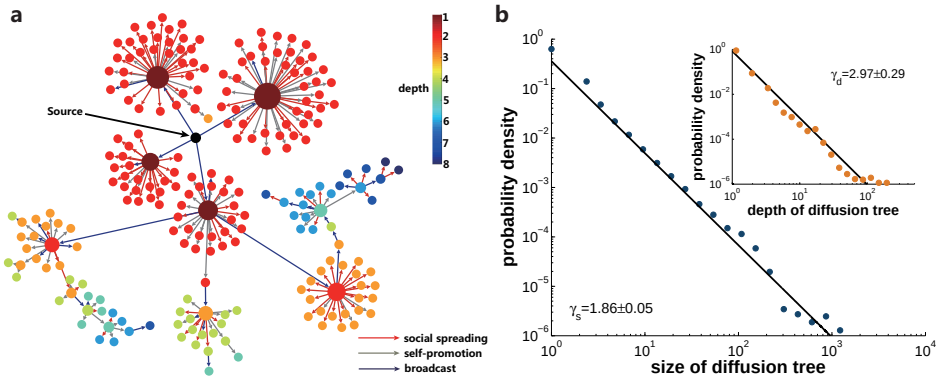

Figure 1. The diffusion trees in LJ community. a, a real instance of information diffusion. An illustration of a diffusion tree containing 227 nodes which reaches the depth of 8. Each node represents a post published in LJ community, whereas each link stands for a spreading instance. The node color indicates the depth of a node in the diffusion tree. The size of a node is proportional to the number of its children. The links of social spreading, self-promotion, and broadcast are represented by the colors of red, grey, and blue respectively. **b** shows the probability distributions of diffusion trees' size and depth. Both the tree size and depth exhibit approximately pow-law distributions. The power-law exponents for tree size and depth are $\gamma_s = 1.86 \pm 0.05$ and $\gamma_d = 2.97 \pm 0.29$ respectively. The straight lines represent the maximum likelihood fitting of the data points.

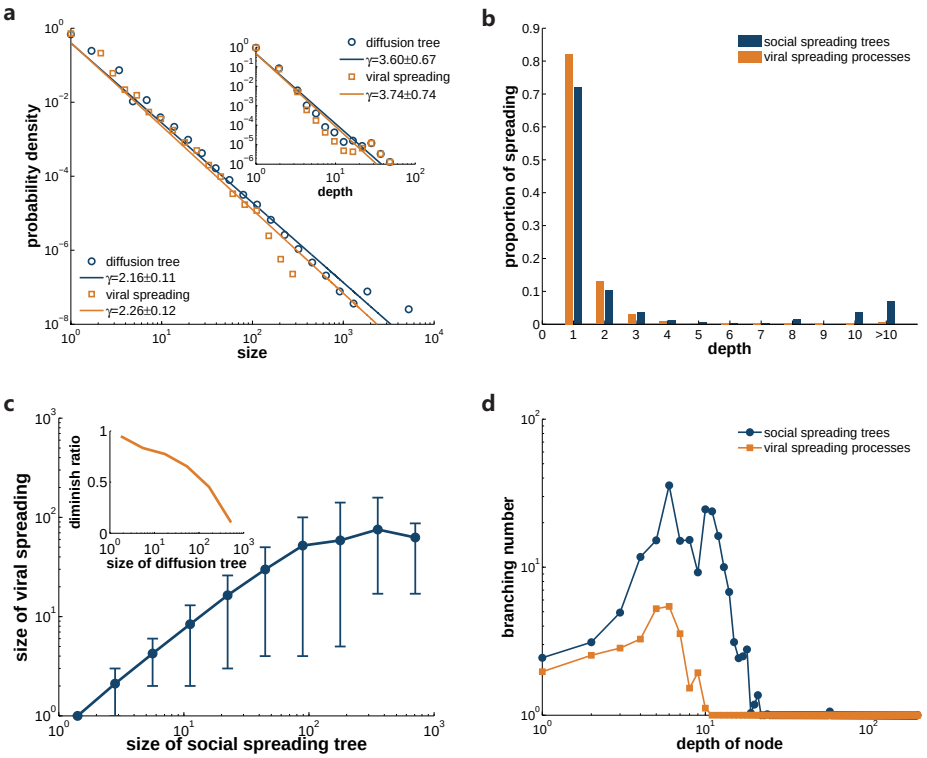

Figure 2. Analysis of social spreading. **a** shows the probability distributions of the size of diffusion trees and viral spreading processes. The inset displays the distributions of spreading depth for both cases. The straight lines are fitted with the maximum likelihood method. In **b**, we present the proportion of diffusion instances in spreading processes with a given depth. The relation between the size of viral spreading and diffusion trees is displayed in **c**. Error bars indicate the 10% and 90% percentiles. The inset presents the diminishing ratio when mapping the diffusion trees to viral spreading. In **d**, we classify the nodes according to their depth in spreading processes and display their average branching number.

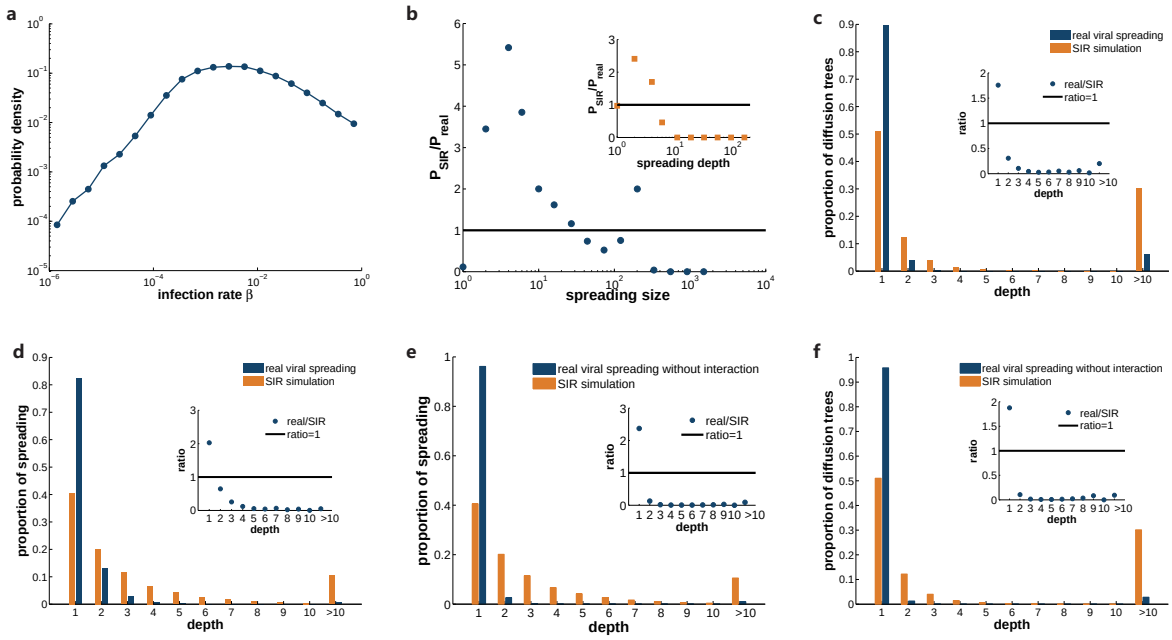

Figure 3. SIR modeling with users' infection rate cannot reproduce the realistic viral spreading pattern. a, distribution of the real-world infection rate for each individual β calculated from viral spreading instances. We display the ratio between the size distribution of SIR simulations and real viral spreading in **b**. Inset shows the ratio of depth distribution. In **c**, we present the proportion of diffusion trees with a given depth for both SIR simulations and real viral spreading, and show the ratio between real cases and SIR modeling in the inset. **d** presents the proportion of spreading instances for diffusion with a specific depth for both cases. The inset shows the ratio between real viral spreading and simulations. In **e** and **f**, we perform same analyses for viral spreading without interactions with other diffusion types.

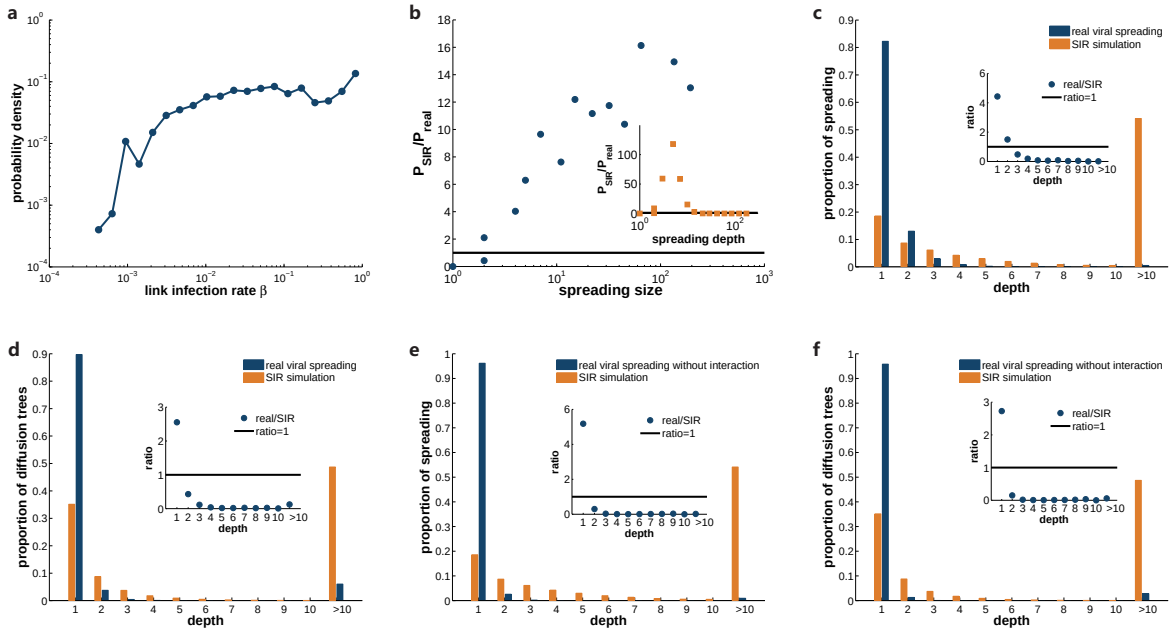

Figure 4. SIR modeling with links' infection rate cannot reproduce the realistic viral spreading pattern. a, distribution of the real-world infection rate for each social link β calculated from viral spreading instances. The ratio between the size distribution of SIR simulations and real viral spreading is displayed in **b**. Inset shows the ratio of depth distribution. In **c**, the proportion of diffusion trees with a given depth for both SIR simulations and real viral spreading is presented, and the ratio between real cases and SIR modeling is shown in the inset. **d** illustrates the proportion of spreading instances for diffusion with a given depth for both cases. The inset shows the ratio between real viral spreading and simulations. Same analyses are shown in **e** and **f** for real viral spreading without interactions with self-promotion and broadcast diffusion.

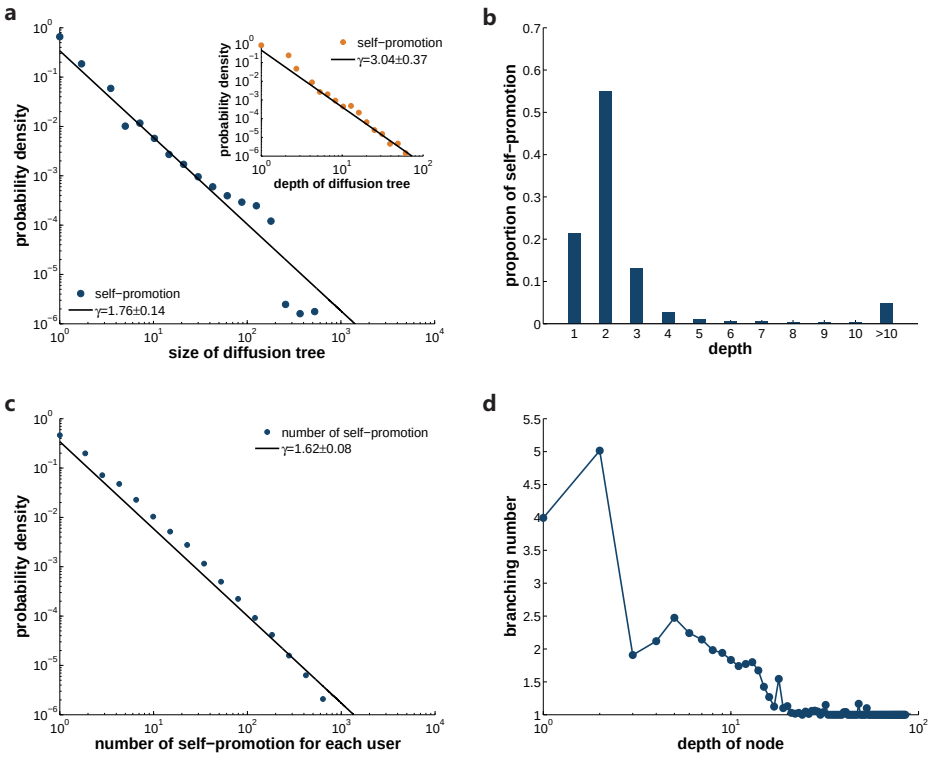

Figure 5. Analysis of the self-promotion. **a** shows the distributions of the size and depth of self-promotion diffusion trees. The fraction of self-promotion links in diffusion trees with a certain depth is displayed in **b**. In **c** we present the probability distribution of the total number of self-promotion for each user, which has a power-law shape with exponent $\gamma = 1.62 \pm 0.08$. In **d** we plot the relationship between posts' branching number and their depth in self-promotion diffusion trees.

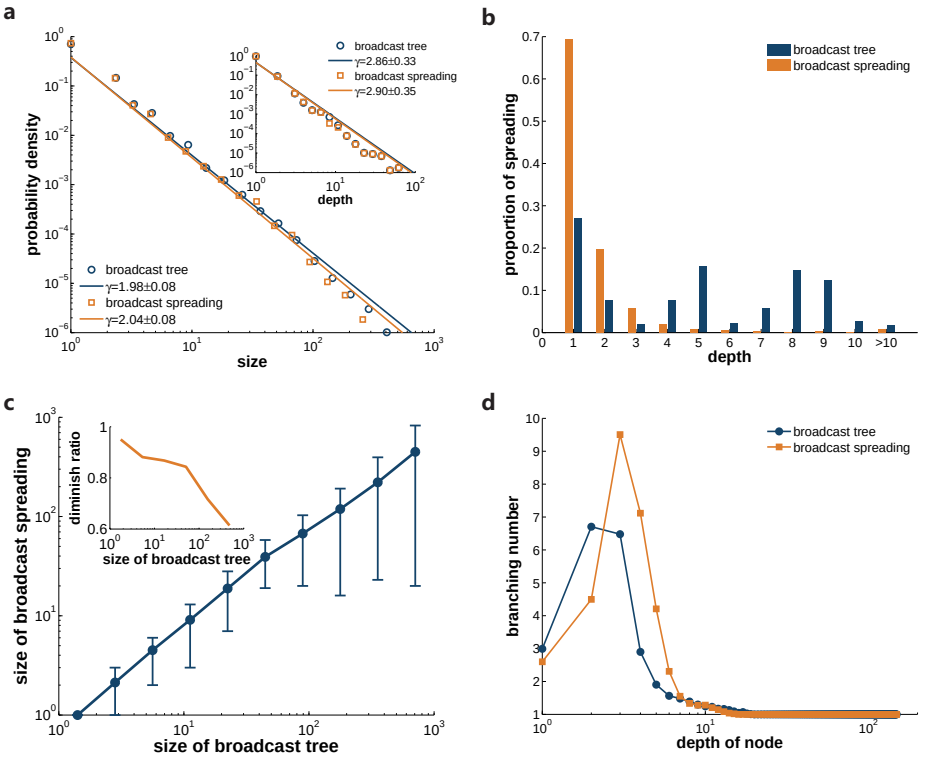

Figure 6. Properties of the broadcast. In **a**, we display the distributions of the size and depth for broadcast diffusion trees and broadcast spreading respectively. The proportion of broadcast links in diffusion processes with a certain depth is shown in **b**. The relation between the size of broadcast spreading and broadcast diffusion trees is displayed in **c**. Error bars indicate 10% and 90% percentiles. The inset presents the diminishing ratio when mapping the diffusion trees to broadcast spreading. We plot nodes' average branching number versus their depth in diffusion in **d**.

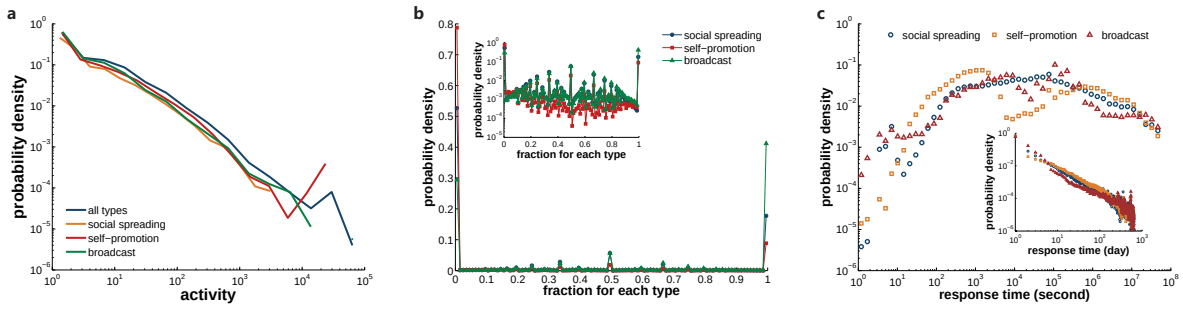

Figure 7. The human activity of LJ users. In **a** we show the distribution of users' activity for each mechanism. For each user, we calculate the fraction of each mechanism that the selected user has conducted, and present the distribution of obtained fraction in **b**. In the inset we change the linear scale of y axis to a logarithmic scale. **c** displays the distribution of response time τ for social spreading, self-promotion, and broadcast. We adopt the time unit of second in the main panel and day in the inset.

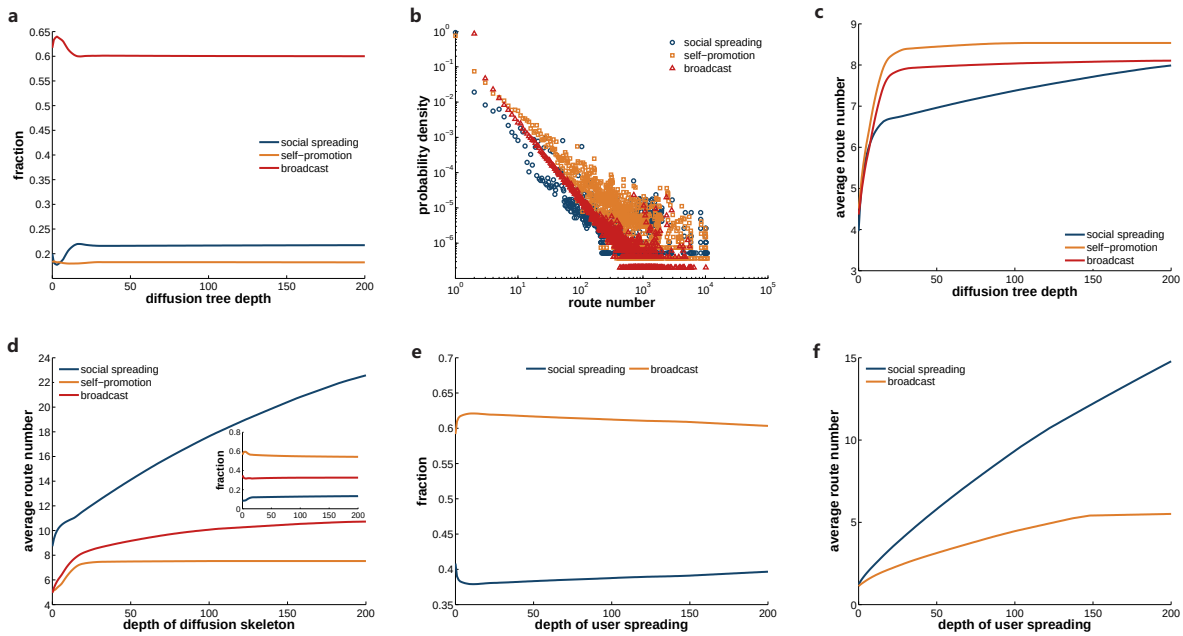

Figure 8. Coupling of distinct mechanisms. We plot the fraction of the social spreading, self-promotion, and broadcast links in diffusion trees deeper than a given threshold in **a**. The x-axis value is the lower bound of selected trees' depth. In **b**, the distribution of the route number for each type is displayed. We calculate the average route number of the diffusion links for each type in diffusion trees deeper than a certain depth, and present the results in **c**. After removing the leaves of diffusion trees, we obtain the information diffusion skeletons. We show the average route number and composition in diffusion skeletons whose depth exceeds certain values in the main panel and inset of **d** respectively. In the spreading processes among population, the fraction and average route number of social spreading and broadcast links are presented in **e** and **f**.

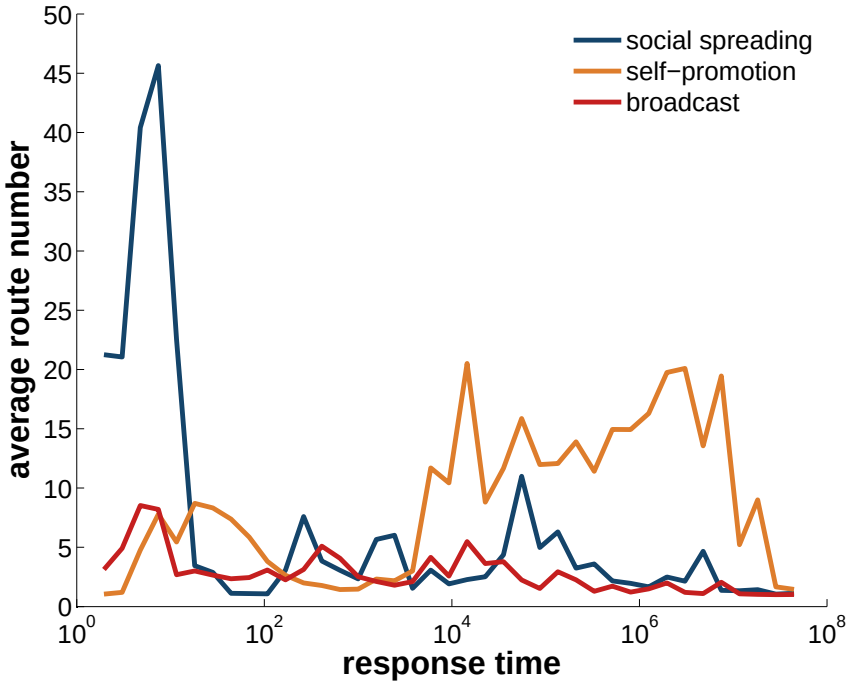

Figure 9. The average route number versus response time for different mechanisms. For each diffusion type, we classify diffusion links according to their response time (in second), or equivalently the diffusion speed, and display the average route number for each case.