ARTICLES
https://doi.org/10.1038/s41562-020-00945-1

nature human behaviour

Realistic modelling of information spread using peer-to-peer diffusion patterns

Bin Zhou[1,2,9]✉, Sen Pei[3,9]✉, Lev Muchnik[4,5], Xiangyi Meng[2], Xiaoke Xu[6], Alon Sela[7], Shlomo Havlin[2,8] and H. Eugene Stanley[2]

In computational social science, epidemic-inspired spread models have been widely used to simulate information diffusion. However, recent empirical studies suggest that simple epidemic-like models typically fail to generate the structure of real-world diffusion trees. Such discrepancy calls for a better understanding of how information spreads from person to person in real-world social networks. Here, we analyse comprehensive diffusion records and associated social networks in three distinct online social platforms. We find that the diffusion probability along a social tie follows a power-law relationship with the numbers of disseminator's followers and receiver's followees. To develop a more realistic model of information diffusion, we incorporate this finding together with a heterogeneous response time into a cascade model. After adjusting for observational bias, the proposed model reproduces key structural features of real-world diffusion trees across the three platforms. Our finding provides a practical approach to designing more realistic generative models of information diffusion.

As a ubiquitous process in social networks, information diffusion plays a central role in applications ranging from the spread of news and opinions[1], the propagation of innovations[2], word-of-mouth viral marketing[3] and change in behaviour[4] to product adoption[5]. A growing number of studies have revealed that information diffusion is a complex process shaped by such interacting factors as network structure[6], information content[7], human activity[8], stochastic dynamics[9] and even users' behavioural[10] and sociodemographic characteristics[11]. These entwined factors give rise to the diverse structure of diffusion paths observed in the real world[12–17], which obscures our understanding of the mechanisms driving spreading dynamics.

To understand how different factors affect information diffusion, the ideal method is to design randomized controlled experiments in real-world social networks[4,18–21]. However, due to the difficulty and cost in implementing such in vivo experiments, in silico agent-based modelling in structured networks has been routinely employed as an alternative to simulate information spread[22–26]. Simplified epidemic-like models have been frequently used and applied in a wide range of studies such as influential spreader identification[27–33], social recommendation[1,3] and rumour containment[34]. In analogy to transmission of contagious diseases[35,36], information in epidemic-like models diffuses along social ties from person to person, with diffusion events independent from each other. Following this rule, consecutive peer-to-peer diffusion can form large-scale deep information cascades lasting multiple generations. As such, these models are also referred to as independent cascade models in literature[5].

Despite the widespread use of such epidemic-like models, the mechanism of biological contagion and information diffusion are fundamentally different. In contrast to epidemic processes in which exposures to infection result in passive transmission, social contagion is a deliberate action taken by individuals who receive information. Empirical studies have demonstrated that simple generative models inspired by epidemic processes fail to reproduce some key features of the observed diffusion trees[14–17]. For instance, contrary to the high frequency of occurrence of multi-generation cascades produced by epidemic-like models, such events were rarely observed in a range of online social platforms[15–17]. This critical discrepancy between observed and model-generated diffusion patterns indicates that a better understanding of how information spreads from person to person is required for the development of more realistic diffusion models.

Several generative models have been developed to reproduce the characteristics of real-world diffusion trees in different settings[13–15,37]. For instance, a probabilistic model based on network clustering and asynchronous response time can generate the narrow and deep tree-like structure of the propagation of Internet chain letters[14]; a branching process incorporated with high variability of human behaviour can replicate realistic distribution of cascade sizes in viral marketing campaigns[37] and a susceptible-infected-recovered (SIR) model in scale-free networks with appropriate parameter settings can reproduce the distribution of structural virality (that is, the average distance between all pairs of nodes in a diffusion tree) for diffusion trees in Twitter[15]. Although these models successfully generated certain features of realistic diffusion trees, due to the unavailability of underlying social connections, most of them were not tested on the same social networks in which the observed diffusion occurred. In particular, these models were unable to relate the propensity of a social tie to convey information to characteristics of the involved users and had to model the probability of diffusion between two users as either a constant or a random number drawn from a predefined distribution. This intuitive setting, simple as it is, has rarely been verified by empirical evidence. Even the literature

[1]School of Economics and Management, Jiangsu University of Science and Technology, Zhenjiang, China. [2]Center for Polymer Studies and Department of Physics, Boston University, Boston, MA, USA. [3]Department of Environmental Health Sciences, Mailman School of Public Health, Columbia University, New York, NY, USA. [4]School of Business Administration, The Hebrew University of Jerusalem, Jerusalem, Israel. [5]Microsoft Research Israel, Alan Turing 3, Hertzliya, Israel. [6]College of Information and Communication Engineering, Dalian Minzu University, Dalian, China. [7]Industrial Engineering Department, Ariel University, Ariel, Israel. [8]Department of Physics, Bar-Ilan University, Ramat Gan, Israel. [9]These authors contributed equally: Bin Zhou, Sen Pei.
✉e-mail: binzhou@mail.ustc.edu.cn; sp3449@cumc.columbia.edu

on complex contagion[38] that studied the impact of the number of exposures on behaviour adoption does not reveal how information diffusion in a structurally heterogeneous social network depends on network attributes.

In this work, we explore the patterns of peer-to-peer information diffusion in online social media, and aim to use these uncovered patterns to develop a generative model capable to better reproduce the observed diffusion trees. Specifically, we focus on the relationship between the diffusion probability along a social tie and the numbers of connections (that is, degrees) of both the disseminator (also commonly named the spreader) and the receiver. To perform this analysis, detailed data of individual-level diffusion events and the associated social network structure are needed. In this study, we use datasets containing both real-world information flows and the underlying social relationships in three distinct online social platforms: a blog-sharing community—LiveJournal (https://www.livejournal.com), and two microblog services—Weibo (https://www.weibo.com) and Twitter (https://twitter.com). These detailed and large datasets allow us to examine the impact of social network structure on diffusion dynamics across diverse social platforms.

Results

Information diffusion in real-world social platforms. LiveJournal is one of the earliest online platforms designed to host and disseminate personal posts. Every LiveJournal user maintains a friend list. The undirected friend relationship is reciprocal, and the entire friendship network, consisting of about 9 million users and 188 million social ties, reflects the social relations among LiveJournal users. In LiveJournal community, users can get access to the posts published by their friends, and may refer to these posts in their own articles by including hyperlinks to the original posts. We therefore use the hyperlink reference to track the information flow from user to user. A total of 721,547 diffusion events following friend relationships during 14 February 2010 to 21 November 2011 were collected. In addition, the total number of posts published by each user during the same period was recorded.

In contrast to LiveJournal in which social relationships are undirected, Weibo and Twitter maintain directed social ties using a 'follow' mechanism: one can follow other users without their consent. Particularly, users can get updates on the 'tweets' posted by their friends (that is, users they follow, or termed 'followees'). Diffusion events between users are identified by the 'retweet' marker. The Weibo dataset contains 9,019,288 diffusion events, involving 4,483,515 users. In addition, we analyse 43,099 retweets in Twitter about the discovery of Higgs boson among 67,680 users[39]. The social networks of Weibo and Twitter, consisting of the friends and followers of the users involved in retweeting, contain about 8 million and 457 thousand users, respectively (see details in Methods, Supplementary Tables 1 and 2 and Supplementary Fig. 1).

Observed peer-to-peer information diffusion. We focus on the local structure of ego-networks formed by connections between users and their neighbours. For LiveJournal, these neighbours are the friends in one's friend list; for Weibo and Twitter, these neighbours are users' followers who can view their tweets. Once a disseminator posts an article or tweet, all neighbours in the ego-network get exposed to the information and become receivers who may further repost it to their neighbours (Supplementary Fig. 2). In real-world diffusion, most receivers exposed to content chose not to repost it[15–17]. Receivers might not repost the content because they saw it but decided not to share, or sometimes they might not even see the content due to overload. Receivers who do decide to share the information with their followers are referred to as adopters. Although receivers who are not adopters do not contribute to information dissemination, they are crucial to assess the diffusion probability between users and to understand the dynamics of underlying information spread processes.

We start by examining the probability of diffusion along a social tie connecting a disseminator with a degree k_d to a receiver with a degree k_r. For the LiveJournal platform, k_d and k_r are the numbers of friends in the undirected social network. For the directed social networks of Weibo and Twitter, k_d represents the number of followers of the disseminator (that is, in-degree) and k_r stands for the number of friends the receiver follows (that is, out-degree). In both cases, k_d is the number of receivers reached by the information, and k_r is the number of potential information sources of the receiver. We group the links from the disseminators involved in the diffusion event to their receivers into a group R. The links in R encompass all observed diffusion 'attempts' of which only a subgroup is successful. Another set S contains only links actively involved in diffusion cascades (that is, successful diffusion paths). Within the groups of R and S, we bin the links with a certain combination of k_d and k_r into subgroups R_{k_d,k_r} and S_{k_d,k_r}. As k_d and k_r are highly heterogeneous, the data bin is performed in the logarithmic scale, with ten bins in each dimension. Aggregation of links into bins is necessary to reduce the number of (k_d,k_r) combinations. Group size $|S_{k_d,k_r}|$ and $|R_{k_d,k_r}|$ represent the numbers of links that match criterion (k_d,k_r) in S_{k_d,k_r} and R_{k_d,k_r} (see Supplementary Fig. 3 for plots of $|S_{k_d,k_r}|$ and $|R_{k_d,k_r}|$). Next, we define the average diffusion probability along social ties with k_d and k_r users as $\Lambda_{k_d,k_r} = |S_{k_d,k_r}|/|R_{k_d,k_r}|$. The quantity Λ_{k_d,k_r} represents the average fraction of the (k_d,k_r) links exposed to the content that were actively involved in the observed information diffusion cascade (Fig. 1a–c).

The stripe pattern in the double-logarithmic scale of Fig. 1a–c suggests a power-law functional form of the diffusion probability $\Lambda_{k_d,k_r} = ck_d^{\alpha}k_r^{\beta}$. We fit the equation parameters using the 10×10 data points in Fig. 1a–c. Specifically, we performed a linear regression to the function $\log\Lambda_{k_d,k_r} = \alpha\log k_d + \beta\log k_r + \log c$ (see Supplementary Table 3 for the fitted parameters). The power-law function explains the data well, as demonstrated in Fig. 1d–f where the data points $k_d^{\alpha}k_r^{\beta}$ are proportional to Λ_{k_d,k_r}.

Adjusting observational bias. Despite the good fitting, the data reported in Fig. 1a–c are in fact subject to an observational bias towards successful diffusion events and overlook unsuccessful diffusion attempts. This observational bias is a major obstacle preventing realistic modelling of information diffusion using passively observed data. To develop a more realistic cascade model for information diffusion, this observational bias needs to be corrected.

Among the examined datasets, only the LiveJournal data contain the number of posts published by each user, including those that were not reposted. To keep the analysis consistent across three platforms, in Fig. 1a, we show the diffusion probability in LiveJournal calculated using only the succeeded diffusion attempts, same as the analysis for Weibo and Twitter. However, given that we have access to all diffusion attempts in LiveJournal, the actual diffusion probability can be obtained by including all attempts into the denominator $|R_{k_d,k_r}|$, which corrects the observational bias. For Weibo and Twitter, such information is unavailable. We therefore need an alternative approximation method to adjust the observational bias in the Weibo and Twitter datasets.

Here we propose a framework to adjust this bias using the Bayes' rule. Define a diffusion attempt from a disseminator with a degree k_d to a receiver with a degree k_r as a (k_d,k_r) attempt. For each (k_d,k_r) attempt, we define the event 'spread' as a successful transmission and the event 'observed' as the attempt being observed (that is, included in the group of diffusion attempts R). Note that for Twitter and Weibo, a (k_d,k_r) attempt is observed if and only if the information is reposted by at least one receiver. Because we only considered diffusion attempts of observed information in previous calculation, the diffusion probability Λ_{k_d,k_r} is actually a conditional probability

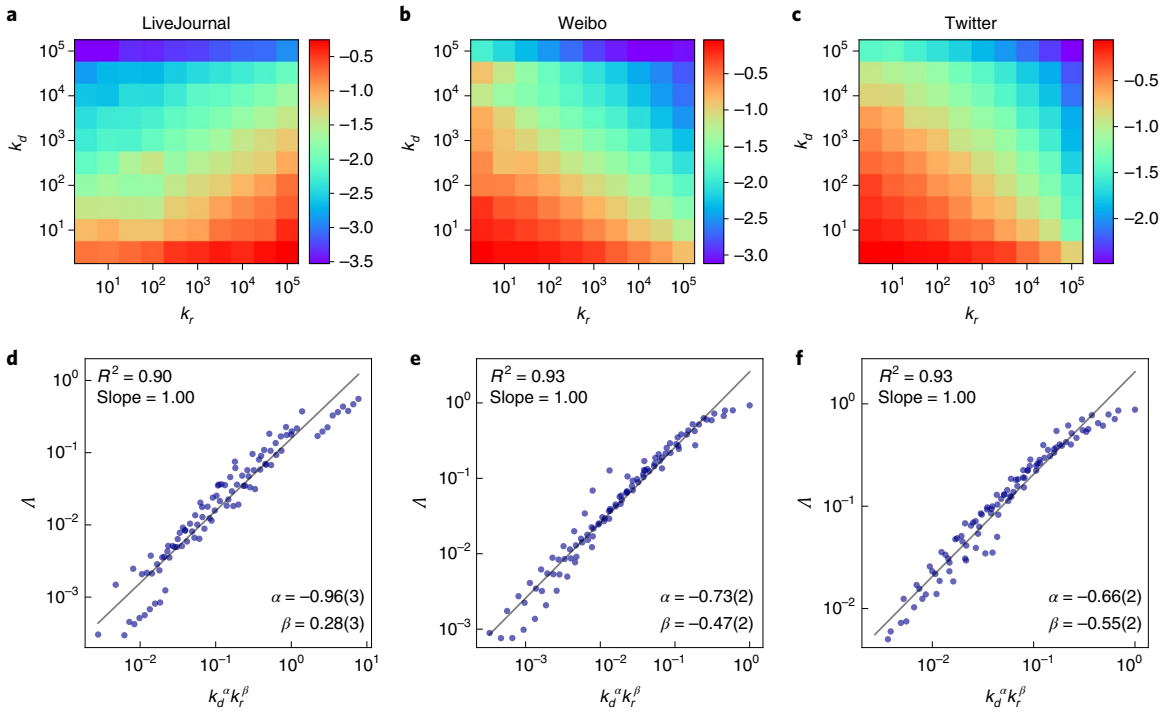

Fig. 1 | The unadjusted diffusion probability Λ_{k_d,k_r} following (k_d,k_r) links. a–c, For LiveJournal (**a**), Weibo (**b**) and Twitter (**c**) datasets, the diffusion probability is calculated as the fraction of successful diffusion events among observed diffusion attempts along social ties from a disseminator with a degree k_d to a receiver with a degree k_r. Colour indicates the logarithmic value of Λ_{k_d,k_r} (base 10). The relationships between Λ_{k_d,k_r} and $k_d^\alpha k_r^\beta$ are shown for LiveJournal (**d**), Weibo (**e**) and Twitter (**f**). Standard errors for α and β are reported in the parentheses. Each dot represents one (k_d,k_r) combination and its corresponding diffusion probability Λ_{k_d,k_r} in **a–c**. Analyses are based on 721,547 diffusion events in LiveJournal, 9,019,288 diffusion events in Weibo and 43,099 diffusion events in Twitter.

$P_{k_d,k_r}(\text{spread}|\text{observed})$—the probability of successful diffusion among the (k_d,k_r) attempts of observed information. In model simulation, however, what we need is $P_{k_d,k_r}(\text{spread})$—the probability of successful diffusion among the (k_d,k_r) attempts of any posted information. Using the Bayes' rule, we have

$$P_{k_d,k_r}(\text{spread}) = \frac{P_{k_d,k_r}(\text{observed})}{P_{k_d,k_r}(\text{observed}|\text{spread})} P_{k_d,k_r}(\text{spread}|\text{observed}),$$

where $P_{k_d,k_r}(\text{observed}|\text{spread})$ is the probability of a (k_d,k_r) diffusion attempt being included in R given the information successfully diffuses along the (k_d,k_r) link, and $P_{k_d,k_r}(\text{observed})$ is the probability of a (k_d,k_r) diffusion attempt being included into R_{k_d,k_r}, no matter whether the information is reposted or not. According to the definition of R_{k_d,k_r}, if a (k_d,k_r) diffusion attempt succeeds, it would be definitely included in R_{k_d,k_r}. As a result, we have $P_{k_d,k_r}(\text{observed}|\text{spread}) = 1$, which leads to

$$P_{k_d,k_r}(\text{spread}) = P_{k_d,k_r}(\text{observed})\Lambda_{k_d,k_r}.$$

This relationship indicates that we can adjust the bias of the conditional probability Λ_{k_d,k_r} by multiplying a factor $P_{k_d,k_r}(\text{observed})$.

To calculate $P_{k_d,k_r}(\text{observed})$, the total number of articles/tweets posted by each user is needed. This information is known for LiveJournal but unavailable for Weibo and Twitter. For LiveJournal, we group the diffusion attempts of all posted information, whether reposted by other users or not, into an adjusted group of attempts, denoted by R^a. We calculate $P_{k_d,k_r}(\text{observed})$ as the fraction of the (k_d,k_r) attempts in R^a that were actually reposted and observed in the group R: $P_{k_d,k_r}(\text{observed}) = \left|R_{k_d,k_r}\right| / \left|R^a_{k_d,k_r}\right|$ (see values of $P_{k_d,k_r}(\text{observed})$ in Supplementary Fig. 4a). The adjusted diffusion probability of LiveJournal is calculated by $\Lambda^a_{k_d,k_r} = P_{k_d,k_r}(\text{observed})\Lambda_{k_d,k_r} = \left|S_{k_d,k_r}\right| / \left|R^a_{k_d,k_r}\right|$.

For Weibo and Twitter, the number of actual diffusion attempts is unknown. In principle, the factor $P_{k_d,k_r}(\text{observed})$ cannot be calculated without this information. However, if the frequency of posting activity for users with a given degree k_d is not extremely heterogeneous, we can approximate $P_{k_d,k_r}(\text{observed})$ as $P_{k_d,k_r}(\text{observed}) \approx \left|O_{k_d,k_r}\right| / \left|G_{k_d,k_r}\right|$, where $\left|G_{k_d,k_r}\right|$ is the number of (k_d,k_r) links in the social network and $\left|O_{k_d,k_r}\right|$ is the number of unique (k_d,k_r) links in R_{k_d,k_r} (see Methods for details on the definitions and calculations, Supplementary Fig. 4b,c). The adjusted diffusion probability is estimated by $\Lambda^a_{k_d,k_r} = \Lambda_{k_d,k_r}\left|O_{k_d,k_r}\right| / \left|G_{k_d,k_r}\right|$. To better illustrate this observational bias correction procedure, we introduce a concrete example in Supplementary Fig. 5.

We note that this approximation method only provides an upper bound for the adjustment factor $P_{k_d,k_r}(\text{observed})$. In our following analysis, we performed a consistency check using the LiveJournal dataset, in which all published posts are available. Results obtained using all posts and the approximation method generally agree with each other. This consistency check provides additional credibility for our analysis on Weibo and Twitter datasets.

Alternative to the approximation method, in real-world applications, the adjustment factor $P_{k_d,k_r}(\text{observed})$ could also be estimated by actively monitoring the activity of a sufficient number of sample users. All posts published by these users could be collected to better estimate the adjustment factor through $P_{k_d,k_r}(\text{observed}) = \left|R_{k_d,k_r}\right| / \left|R^a_{k_d,k_r}\right|$, similar to our analysis on the LiveJournal dataset.

Patterns of peer-to-peer information diffusion. We apply the method proposed in the last section to adjust the observational bias in the results reported in Fig. 1a–c. In general, after adjusting

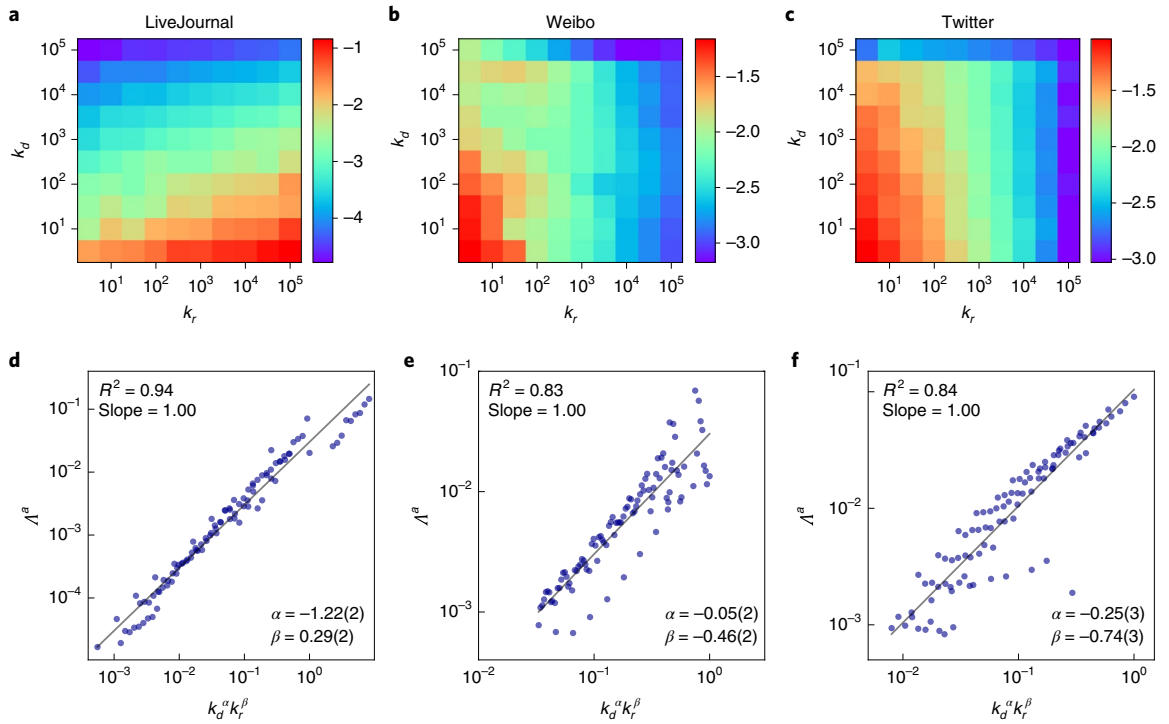

Fig. 2 | The power-law relationship between peer-to-peer diffusion probability and users' degrees. a–c, The adjusted peer-to-peer diffusion probability $\Lambda^a_{k_d,k_r}$ following (k_d,k_r) links for LiveJournal, Weibo and Twitter. The logarithmic values of Λ_{k_d,k_r} (base 10) are represented by colours. While Weibo (**b**) and Twitter (**c**) share similar peer-to-peer diffusion pattern, LiveJournal (**a**) exhibits fundamentally different behaviour. **d–f**, The relationships between the adjusted diffusion probability Λ^a and $k_d^\alpha k_r^\beta$ for LiveJournal (**d**), Weibo (**e**) and Twitter (**f**). The fitted parameters α and β are reported with standard errors in the parenthesis. Analyses are based on 721,547 diffusion events in LiveJournal, 9,019,288 diffusion events in Weibo and 43,099 diffusion events in Twitter.

the observational bias, the diffusion probability $\Lambda^a_{k_d,k_r}$ follows a power-law relationship with the disseminator and recipient's degree: $\Lambda^a_{k_d,k_r} = ck_d^\alpha k_r^\beta$. We used the data for a 10×10 grid of k_d and k_r values in Fig. 2a–c, and performed a linear regression to the function $\log\Lambda^a_{k_d,k_r} = \alpha\log k_d + \beta\log k_r + \log c$. The fitted exponents are reported in Fig. 2d–f and Supplementary Table 4. Consistent with Fig. 2a–c, the exponent α is negative for all platforms; β is negative for Weibo and Twitter, but positive for LiveJournal. Figure 2d–f indicates that $\Lambda^a_{k_d,k_r}$ is proportional to $k_d^\alpha k_r^\beta$.

We find that dependence of the adjusted diffusion probability $\Lambda^a_{k_d,k_r}$ on k_d is universal across all three platforms: the information posted by a highly connected disseminator is in fact less likely to be reposted by each of the followers (Fig. 2a–c). In other words, effectiveness of the social network actors to spread information drops with their degree. We speculate that this counterintuitive result could be explained by the fact that recipients of content from highly connected individuals could perceive the information published by those 'hubs' as too widespread (and not sufficiently new)[40]. An alternative structural explanation could be that users connected to hubs are probably linked to multiple hubs active in posting and suffer from information overload, which leads them to respond less to the content they receive[41]. We examined the probability of reposting as a function of the number of posts each user was exposed to, and verified that the reposting probability generally decreases as the number of received information grows (Supplementary Fig. 6).

In contrast with the dependence of $\Lambda^a_{k_d,k_r}$ on k_d, dependence of responsiveness of the users exposed to content on their degree k_r differs for different platforms: it drops with increasing k_r in both microblogging platforms (on the basis of directional networks) (Fig. 2b,c), but increases with k_r in the undirected network of blogs (Fig. 2a). In microblog service, information from a large number of sources may compete for receivers' attention and thus reduce the diffusion probability[42–44]. In particular, several studies have reported similar observations in microblog websites. For instance, it has been found that the finite ability to process incoming information constraints social contagion in Twitter, and the probability of retweeting a URL decays as a power-law function of the number of friends[45,46]. In LiveJournal, however, information is more likely to be reposted by well-connected receivers. This dramatic difference is potentially caused by the distinct behaviour of users in blog communities. Specifically, it is possible that users of a blog-sharing community that uses reciprocal connections such as LiveJournal need to maintain a high frequency of posting to attract friends and achieve a large degree, as observed in other online content-sharing platforms[47]. As a result, well-connected users tend to be more actively involved in reposting in LiveJournal. In Supplementary Fig. 7, we show that the numbers of reposts of LiveJournal users have a stronger positive correlation with their degrees. In addition, the network directionality can reinforce this effect. In LiveJournal, the bidirectional links effectively limit social connections to mutual acquaintances, which is likely to increase the relevance of information. Consequently, active LiveJournal users, typically with high degrees, tend to repost more frequently than Weibo and Twitter users who may suffer information overload and receive less relevant information. We leave identification of the reasons for the difference between the two kinds of platforms to further research.

We remark that the variation in the change of α after bias correction in different platforms does not undermine the validity of the method. Instead, it highlights the radical difference in users' behaviour between different social platforms—a blog community and two microblog services. Such behavioural difference is encoded in the factor P_{k_d,k_r}(observed), which should be uniquely defined for each platform. In this study, we did not attempt to show that all social platforms share the same diffusion pattern, but to develop a

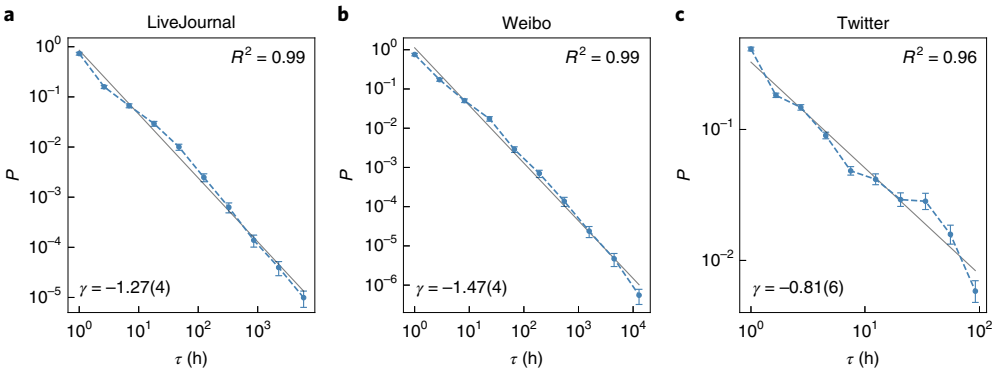

Fig. 3 | The power-law distributions of response time τ. a–c, The distributions of response time for LiveJournal (**a**), Weibo (**b**) and Twitter (**c**), with the fitted power-law exponents reported. The 95% confidence intervals were calculated using bootstrap. Analyses are based on 721,547 diffusion events in LiveJournal, 9,019,288 diffusion events in Weibo and 43,099 diffusion events in Twitter.

generic method to estimate diffusion patterns in social platforms using available data.

Realistic modelling of information diffusion. The same power-law functional form describes peer-to-peer diffusion patterns in all three systems despite considerable differences in platform implementation, social mechanisms and user behaviours. We leverage this to develop more realistic models capable of reproducing the structure of observed diffusion trees. Considering that diffusion of a single piece of information is stochastic and could result in diverse diffusion trees in different model realizations, we test performance of our models by examining generic structural properties of diffusion trees generated from numerous simulations. In particular, we focus on three features: (1) diffusion size N—the total number of users in a diffusion tree; (2) diffusion depth L—the largest number of generations in a cascade and (3) structural virality D—the average distance between all pairs of nodes in a diffusion tree. The structural virality characterizes the high-level structure of diffusion trees: given the same diffusion size, a small D corresponds to a shallow and wide tree, and a large D corresponds to a deep and narrow tree[15]. Across three platforms, the diffusion size and depth follow heavy-tailed distributions, indicating that most observed diffusion cascades are small and shallow. Indeed, the structural virality of diffusion trees is limited to small values. This finding is in agreement with the results in other online social systems[15,16].

We implement a data-driven simulation to reproduce information cascades using parameters inferred from the fitted $\Lambda^a_{k_d,k_r}$ to $k^\alpha_d k^\beta_r$ (Methods and Supplementary Table 4). The simulation realizes the following two mechanisms. First, diffusion probability for each diffusion attempt accounts for the degrees of the disseminator (k_d) and the receiver (k_r) and is calculated through $\Lambda^a_{k_d,k_r} = ck^\alpha_d k^\beta_r$. Second, in compliance with existing literature[8,14,17,37,48,49], response times of the realized diffusion attempts (that is, the time it takes for a receiver to repost the information) are drawn from a power-law distribution $P(\tau) = d\tau^\gamma$ (Fig. 3). Such asynchronous response time has been reported in other social systems, and was used to model real-world information diffusion[14,37]. We confirm the distribution in our data and infer the response time power-law exponents (see Supplementary Table 5). Once a receiver reposts the information, he/she becomes a new disseminator and may trigger further diffusion. The diffusion terminates when no reposting occurs. For comparison, two additional sets of simulations were performed. In the first, we ran a susceptible–infected–recovered model using a constant diffusion probability computed directly from the observed diffusion events ($\Lambda = |S|/|R|$, the mean diffusion probability averaged over all observed attempts; here R is the group of all diffusion attempts and S contains all observed successful diffusion paths); in the second, we ran a cascading model using the unadjusted diffusion probability $\Lambda_{k_d,k_r} = ck^\alpha_d k^\beta_r$ (Fig. 1 and Supplementary Table 3). For each set of simulations, we used the distribution of response time fitted specifically to the diffusion data collected from that platform.

Figure 4 shows distributions of diffusion size N, depth L and structural virality D generated by the simulations described before. Cascades produced by SIR models and simulations based on unadjusted diffusion probability Λ_{k_d,k_r} are substantially larger than the actually observed ones. The adjustment effectively reduces the discrepancy between the observed and model-generated distributions for all features. The distributions obtained from the adjusted model agree better with the observed distributions for all three systems. Although the examined social platforms differ considerably in their social mechanisms, the adjusted models provide a generic generative method to simulate online information diffusion. This close match of distributions also indicates that the approximation method for computing P_{k_d,k_r}(observed) in Weibo and Twitter is effective. Certain discrepancies between the simulated and observed distributions of diffusion size, depth and structural virality still exist, which may be attributed to more complex factors not considered in the models.

As the structural features of diffusion trees are obtained after diffusion terminates, response time does not affect the distributions of diffusion tree size, depth and structural virality. To further explore the impact of heterogeneous response time on simulated information diffusion, we performed a comparison on the distribution of the lifetime T of diffusion trees, that is, time (in hours) between the posting of the first and last post/tweet in a diffusion tree. Specifically, we generated diffusion events using both the response time distribution fitted to observations and a constant response time, set as the mean value of the observed response time. Results in Fig. 5 indicate that a homogeneous response time substantially shortens the lifetime of diffusion trees. This analysis highlights the importance of using a realistic response time distribution in simulating information diffusion[49].

In application, it is desirable to estimate the parameters α and β using samples of diffusion events so that the peer-to-peer diffusion pattern can be generalized to model information spread in the same platform. To check the generality of the power-law diffusion pattern, we estimated the adjusted parameters α and β on the basis of 2, 4, 6, 8, 10 and 20% up to 100% of observed diffusion trees in three platforms. Supplementary Fig. 8 shows the estimation results in the occasion where only the sampled diffusion trees were observed. Generally, the estimated parameters are close to the results obtained using all observed diffusion trees. This indicates

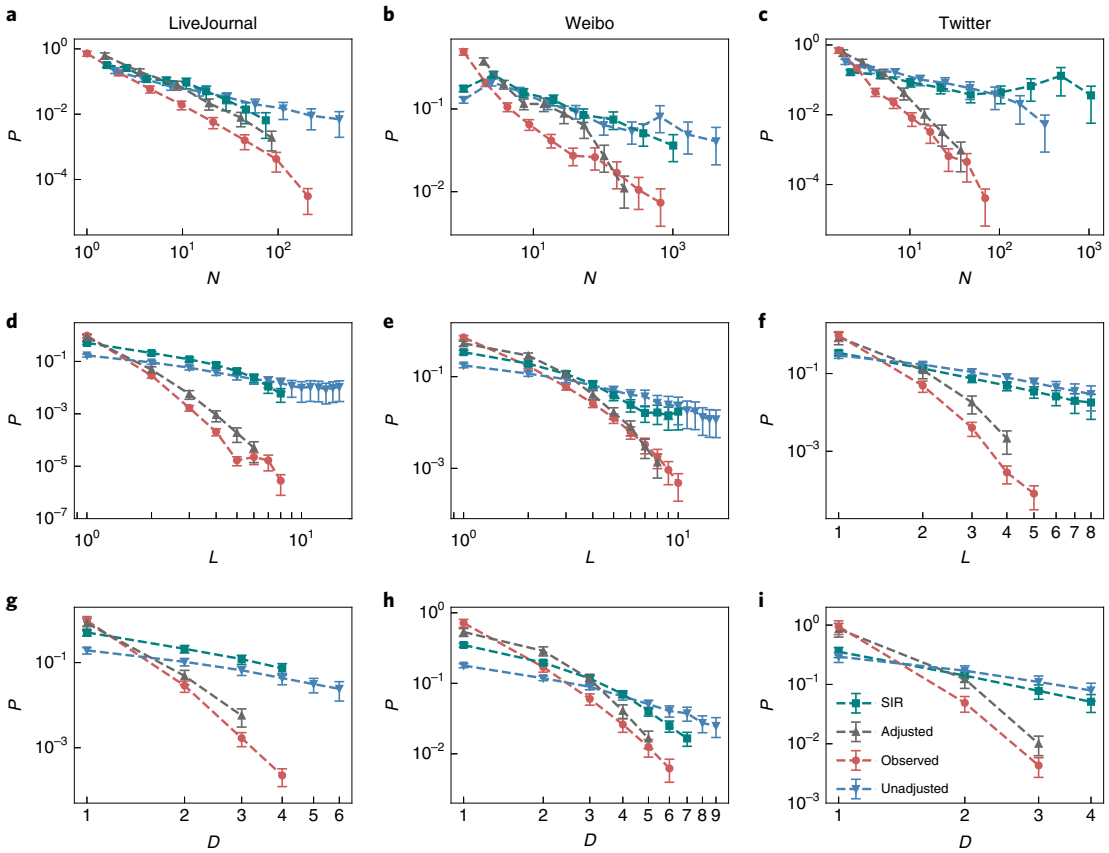

Fig. 4 | Comparison of distributions of diffusion tree size *N*, depth *L* and structural virality *D*. Distributions obtained from the observed data (observed), SIR simulations (SIR), simulations using parameters fitted directly to the data (unadjusted) and simulations using adjusted parameters (adjusted) are distinguished by different symbols. We compare the distributions of diffusion tree size *N*, depth *L* and structural virality *D* for LiveJournal (**a**,**d**,**g**), Weibo (**b**,**e**,**h**) and Twitter (**c**,**f**,**i**). The 95% confidence intervals were obtained by bootstrapping. Analyses are based on 2 million model simulations for each platform.

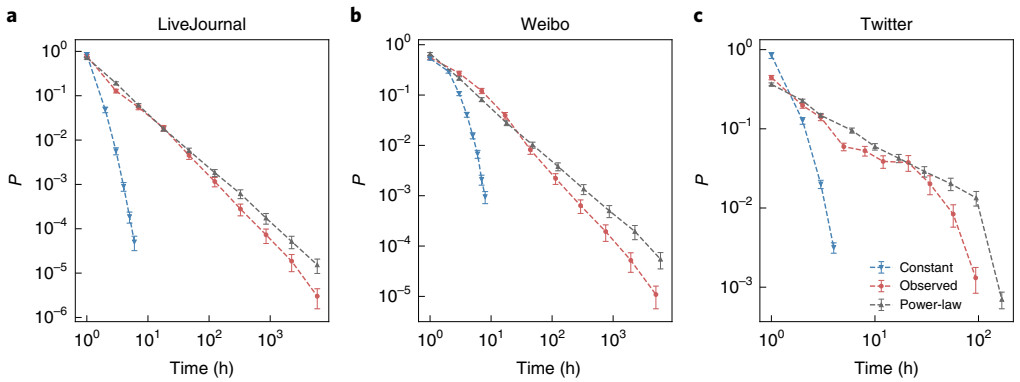

Fig. 5 | Distributions of the lifetime *T* of diffusion trees. Distributions of the lifetime *T* of diffusion trees obtained from the observed data (observed), simulations using a power-law response time (power-law) and simulations using a constant response time (constant) are compared. Comparisons are shown for LiveJournal (**a**), Weibo (**b**) and Twitter (**c**). The 95% confidence intervals were obtained by bootstrapping. Analyses are based on 2 million model simulations for each platform.

that it is feasible to learn the peer-to-peer diffusion pattern by sampling a fraction of diffusion events.

Validation of the observational bias correction using LiveJournal data. To validate the observational bias correction applied to Weibo and Twitter, we performed a consistency check using LiveJournal data, for which all published posts are available. In particular, we estimated the parameters α and β using the actual diffusion attempts as well as the estimated adjustment factor, obtained using the same method applied to the Weibo and Twitter data. The estimated power-law exponents for LiveJournal are $\alpha=-1.22(2)$ and $\beta=0.29(2)$ using the actual attempts, and $\alpha=-0.94(2)$ and $\beta=0.23(2)$ using the approximation method. The diffusion pattern remains the same, and the estimated parameters generally match

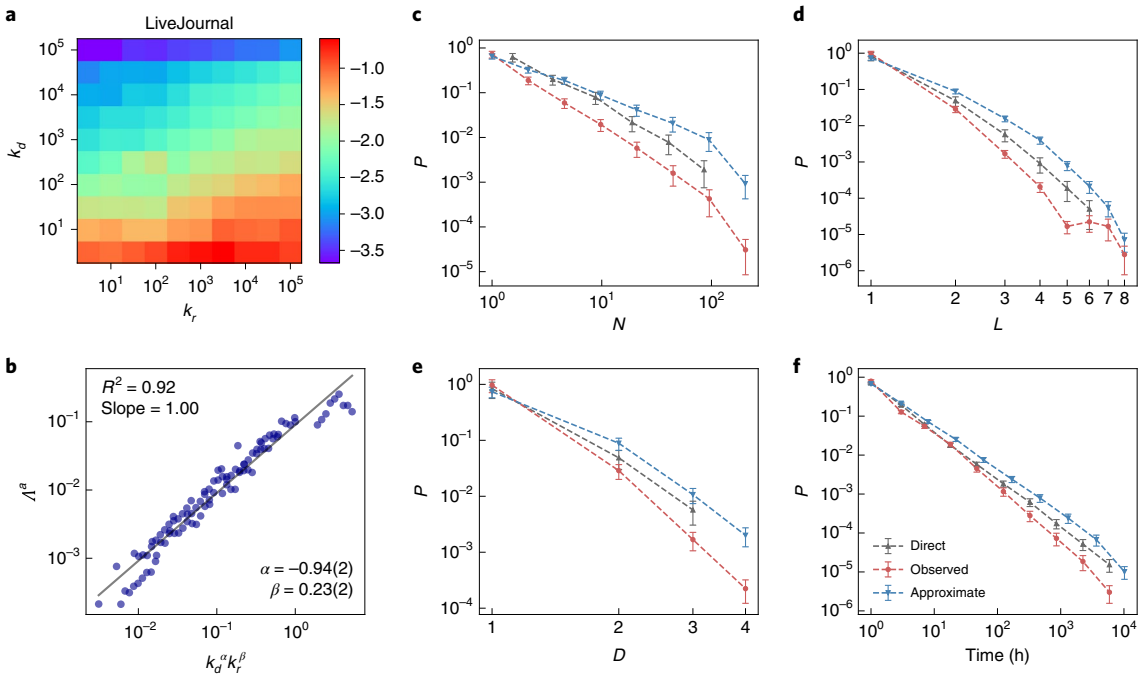

Fig. 6 | A consistency check for the observational bias correction using LiveJournal data. a, The adjusted peer-to-peer diffusion probability for LiveJournal computed using the approximated adjustment factor. **b**, The power-law relationship $\Lambda^a \propto k_d^\alpha k_r^\beta$. **c**–**f**, We compare the distributions of diffusion tree size N (**c**), depth L (**d**), structural virality D (**e**) and lifetime T (**f**) obtained from observed data (observed), simulations based on parameters computed directly from all posts (direct) and simulations based on parameters approximated using the bias correction (approximate). Analyses are based on 721,547 diffusion events in LiveJournal and 2 million model simulations.

their actual values in magnitude. Using the estimated parameters, we reran model simulations and compared the distributions of diffusion size, depth, structural virality and diffusion tree lifetime with those obtained using actual parameters (Fig. 6). For each feature, distributions obtained using actual and estimated parameters generally agree with each other (see the power-law fitting parameters in Supplementary Table 6). Since the estimate of P_{k_d,k_r}(observed) is an upper bound, the diffusion trees generated using the estimated parameters tend to be larger and deeper. This pattern is consistent with what we observed for Weibo and Twitter; that is, the model-generated diffusion trees are generally larger and deeper than observed ones. This consistency check provides additional validation of the observational bias correction method.

Discussion

In this work, we demonstrate that an independent cascade model incorporated with peer-to-peer diffusion patterns and a heterogeneous response time is able to generate key structural features of real-world diffusion trees. Compared with previous modelling works, our approach is more realistic as the model simulations were performed on the same social networks where the diffusion occurs and used parameters learned from the observation data. Given that information diffusion depends on a number of factors, it is surprising that such a simple model can reproduce important statistical properties of realistic diffusion trees. One possible explanation may be that the effect of dominating factors, for example, activity frequency and local social network structure, has been implicitly reflected by the power-law peer-to-peer diffusion patterns.

The specific probability form $ck_d^\alpha k_r^\beta$ is directly derived from diffusion data. For LiveJournal, which we have the full records of posts, this form is the true diffusion pattern. Some other studies have also reported the power-law dependence of diffusion probability on users' degree[45,46]. On the basis of this evidence, the power-law diffusion pattern probably holds in a range of social platforms. The choice of this specific form is not based on assumptions but is supported by empirical diffusion data. Even using a power-law form, the structure of simulated diffusion trees can be far from that of those observed if the parameters are mis-specified. This is evidenced in the simulation using uncorrected parameters α and β. This failure to reproduce the observed diffusion trees using the unadjusted power-law function in turn demonstrates the importance of the observational bias correction. In a realistic model, two ingredients are necessary: (1) a correct form of diffusion probability derived from empirical data and (2) accurate estimation of parameters using observational bias correction. Without either of them, the model may not be able to generate realtistic diffusion trees. We note that, even if a model using a constant diffusion probability could reproduce the observed diffusion, this model is not realistic, as it contradicts the power-law form of diffusion probability observed in empirical data.

This work contributes to the existing literature on information diffusion in two ways. First, we demonstrate that the diffusion probability along a social tie is a product of power-laws of degrees of the disseminator and the receiver. The power-law exponents are different across social platforms and can be inferred from observed diffusion events. Second, we propose a framework to account for the bias in observed diffusion data to develop a more realistic model. We adjust for the bias using the actual number of diffusion attempts for LiveJournal, and provide an approximation method for Weibo and Twitter. Information diffusion cascades generated by the suggested model fit the structural properties of the observed cascades in these systems.

Network abstraction of social systems differs from random mixing because it allows for heterogeneity of exposure. Such models predict a very special role of hubs that ought to have disproportional effect due to the number of exposures they can generate.

Same kinds of models predict that agents having exceptionally high numbers of incoming connections are expected to be exposed to a large volume of information and receive that information early in the diffusion process[50]. In this work we enrich the now-classical network abstraction by adding an additional layer of heterogeneity that dramatically affects dynamics of information cascades. In particular, we demonstrate that the probability to share the information received via social network tie depends on in-degree of the source node as well as the out-degree of the information recipient. The earlier factor could, for example, be considered as a proxy of the information uniqueness: information received from a network hub is less likely to be perceived by its recipient as unique and is less likely to be shared[40]. Simultaneously, social network users who follow many sources experience information overload and have to be very selective about the information they choose to share[42–44]. These mechanisms need further empirical or experimental validations in future works.

In the proposed model, we assume that person-to-person diffusion events are independent and neglect the effect of complex contagion; that is, exposure to information from multiple disseminators simultaneously[4,38]. Nevertheless, as most diffusion events were reposted within a short response time (a few hours), it is reasonable to use an independent cascade model. In a recent experimental study that used Twitter bots[51], it was found that complex contagion models outperform simple contagion models (that is, independent cascade models) in explaining information spread. However, the contagion models used in the comparison are equipped with a constant transmission rate, which is at odds with real-world diffusion. How to disentangle and evaluate the impacts of complex and simple contagion requires more detailed data or controlled experiments.

Our work focuses on diffusion that follows explicit social ties. Currently, we disregard other diffusion mechanisms that do not rely on social relationships such as broadcast and self-promotion[15,17]. Understanding their role in information spread dynamics in online social networks may require construction of hybrid diffusion models driven by multiple mechanisms. Further, quantification and inclusion of more specific features of information, for instance, novelty, validity or virality[52,53], could possibly further improve understanding of information dissemination. In addition, whether similar topics in the same platform exhibit similar peer-to-peer diffusion patterns needs exploration using more comprehensive datasets with text or hashtag information.

Methods

Data. In this study, we used datasets containing both information diffusion records and the associated social networks for three online social platforms—a community of bloggers, LiveJournal, and two microblog services, Weibo and Twitter.

In LiveJournal, each user maintains a friend list representing social ties to other users. The undirected friend relationships form the social networks among 9,573,127 LiveJournal users. Users are notified of the posts published by their friends and may reference these posts in their own articles. These references reveal the diffusion cascades and allow direct observation of the information passed from one user to another. In the dataset, we identified 721,547 diffusion events that follow social ties during 14 February 2010 to 21 November 2011. These diffusion events form 357,749 diffusion trees involving 165,508 users. The number of posts published by each LiveJournal user during the same period was also recorded. In addition, we have collected the underlying social network that contained 188 million friendship relationships. This dataset has been previously used in analysing the pattern of information diffusion[17,54].

In Weibo and Twitter, users keep track of the posts published by the users they follow. This enables information cascades on top of the underlying social network. We infer the person-to-person diffusion events in Weibo and Twitter using 'retweet' mentions. For Weibo, 9,019,288 diffusion events were collected, from which 397,445 diffusion trees containing 4,483,515 unique users were reconstructed. The social network structure, consisting of 7,977,942 users and 700 million following links, was obtained by crawling the friends/followers relationships between the users involved in retweet. The Weibo dataset was released as the training data in an open challenge at https://www.dcjingsai.com/v2/cmptDetail.html?id=166, and is publicly available. The Twitter dataset was collected by monitoring retweets on Twitter about the discovery of Higgs boson around 7 July 2012, when the news was announced. In total, 24,581 retweets diffusion trees were reconstructed from 43,099 diffusion events. The social network that includes the 67,680 users involved in retweeting activities contains 456,626 unique users with 15 million following links. The Twitter dataset was used to explore to diffusion of scientific rumours[39], and is available at https://snap.stanford.edu/data/higgs-twitter.html.

In general, three types of diffusion exist in online social media[17]: social diffusion (that is, A retweets B, A follows B), broadcasting or mediated diffusion (that is, A retweets B, A does not follow B) and self-promotion (that is, A retweets A). In data preprocessing, all retweets that do not follow social links were discarded. Such operation removed broadcasting or mediated diffusion and self-promotion, but guaranteed that all retweets considered in the analysis were between neighbours in social networks.

Estimation of adjustment factor. We assume that users with the same degree k_d have similar posting frequency in Weibo and Twitter. Define the adjustment factor P_{k_d,k_r}(observed) as the probability of a (k_d,k_r) diffusion attempt being observed (that is, present in R_{k_d,k_r}). If each user with a degree k_d posts n tweets during observation, the total number of observed diffusion attempts in R_{k_d,k_r} satisfies $|R_{k_d,k_r}| = nP_{k_d,k_r}(\text{observed})|G_{k_d,k_r}|$, where $|G_{k_d,k_r}|$ is the total number of (k_d,k_r) links in the social network (in the example in Supplementary Fig. 5, the set of all (5,3) links in the social network is $G_{5,3}=\{a\to b, a\to c, a\to e, d\to c, d\to e\}$). To estimate P_{k_d,k_r}(observed), the number of tweets posted by each user, n, is needed. Here, we approximate the lower bound of n using the repetition of links in the group of observed diffusion attempts R_{k_d,k_r}.

Supposing each user is allowed to post only one tweet ($n=1$), the (k_d,k_r) links in R_{k_d,k_r} would have no repetition, as each (k_d,k_r) link has only one chance to be observed. That is, $|R_{k_d,k_r}| = |O_{k_d,k_r}|$, where $|O_{k_d,k_r}|$ is the number of unique (k_d,k_r) links in R_{k_d,k_r}. For $n=2$, each user can perform two rounds of posting. In this case, some (k_d,k_r) attempts may be observed in only one round, thus we have $|R_{k_d,k_r}| \le 2|O_{k_d,k_r}|$. This inequality can be generalized to $|R_{k_d,k_r}| \le n|O_{k_d,k_r}|$ for $n>2$. We estimate the lower bound of n by $n \ge |R_{k_d,k_r}|/|O_{k_d,k_r}|$, which quantifies the average repetition of (k_d,k_r) links in R_{k_d,k_r}. In the example in Supplementary Fig. 5, this average repetition of (5,3) links in our example is $|R_{5,3}|/|O_{5,3}| = 6/3 \le 2 = n$, where $R_{5,3} = \{(a\to b)_1, (a\to c)_1, (a\to e)_1, (a\to b)_2, (a\to c)_2, (a\to e)_2\}$ and $O_{5,3}=\{a\to b, a\to c, a\to e\}$.

For Weibo and Twitter, little information of users' posting activity is available. As a result, here we approximate n using its lower bound: $n \approx |R_{k_d,k_r}|/|O_{k_d,k_r}|$. For disseminators with large k_d, this approximation is more accurate, as their tweets are more likely to be reposted and observed due to a larger number of receivers. The adjustment factor P_{k_d,k_r}(observed) is then estimated using its upper bound: $P_{k_d,k_r}(\text{observed}) = |R_{k_d,k_r}|/(n|G_{k_d,k_r}|) \le |O_{k_d,k_r}|/|G_{k_d,k_r}|$. Since $|O_{k_d,k_r}| \le |G_{k_d,k_r}|$, it is guaranteed that $P_{k_d,k_r}(\text{observed}) \le 1$. Even though P_{k_d,k_r}(observed) is generally overestimated, this adjustment substantially reduces the discrepancy between the distributions of attributes of the observed and model-simulated diffusion trees. Using this approximation, the adjusted diffusion probability is estimated by $\Lambda^a_{k_d,k_r} = \Lambda_{k_d,k_r}|O_{k_d,k_r}|/|G_{k_d,k_r}|$. This approximation turns out to yield satisfactory performance in model simulations. In Supplementary Fig. 4, we report the behaviour of the adjustment factor P_{k_d,k_r}(observed) against k_d and k_r values. We found that P_{k_d,k_r}(observed) also follows a power-law relationship with k_d and k_r: $P_{k_d,k_r}(\text{observed}) = hk_d^{\eta}k_r^{\mu}$ (Supplementary Table 7).

Model simulations. We performed model simulations in real-world social networks. For each social platform, 20,000 observed diffusion trees were randomly selected without replacement. We ran 100 independent simulations of contagion, each started from the root of each diffusion tree. In total, 2 million simulations were performed for each of the three versions of simulations: the SIR model, and cascading models with the unadjusted (Λ_{k_d,k_r}) and adjusted diffusion probability ($\Lambda^a_{k_d,k_r}$). For all three sets of simulations, we used the same power-law response time learned from the observed data. In the SIR model, the diffusion probability was set as the average value computed from the observed diffusion events, that is, $\Lambda = |S|/|R|$. The 95% confidence intervals in Figs. 4 and 5 were obtained by bootstrapping[55]. Specifically, from the simulated diffusion trees, we drew 10^4 random samples consisting of the same number (2 million) of diffusion trees uniformly with replacement, designating them as bootstrap samples. For each bootstrap sample, we calculated the distributions and confidence intervals of diffusion tree size, depth, structural virality and lifetime. The 95% confidence intervals in Figs. 3–6 were computed using the 2.5 and 97.5% percentiles of the 10^4 bootstrap samples of the probability density corresponding to each x axis value.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Weibo and Twitter data are publicly available at https://www.dcjingsai.com/v2/cmptDetail.html?id=166 (in Mandarin) and https://snap.stanford.edu/data/

higgs-twitter.html. LiveJournal data are subject to restrictions for user privacy protection. Interested readers should contact L. Muchnik (lev.muchnik@huji.ac.il) to gain access to the LiveJournal dataset.

Code availability

Custom code that supports the findings of this study is available at https://github.com/bnzu/main-code-of-rmis.

Received: 1 October 2019; Accepted: 6 August 2020;
Published online: 28 August 2020

References

1. Watts, D. J. & Dodds, P. S. Influentials, networks, and public opinion formation. *J. Consum. Res.* **34**, 441–458 (2007).
2. Rogers, E. M. *Diffusion of Innovations* (Simon and Schuster, 2010).
3. Leskovec, J., Adamic, L. A. & Huberman, B. A. The dynamics of viral marketing. *ACM Trans. Web* **1**, 5 (2007).
4. Centola, D. The spread of behavior in an online social network experiment. *Science* **329**, 1194–1197 (2010).
5. Aral, S. & Walker, D. Creating social contagion through viral product design: a randomized trial of peer influence in networks. *Manag. Sci.* **57**, 1623–1639 (2011).
6. Newman, M., Barabási, A. L. & Watts, D. J. *The Structure and Dynamics of Networks* (Princeton Univ. Press, 2011).
7. Vosoughi, S., Roy, D. & Aral, S. The spread of true and false news online. *Science* **359**, 1146–1151 (2018).
8. Iribarren, J. L. & Moro, E. Impact of human activity patterns on the dynamics of information diffusion. *Phys. Rev. Lett.* **103**, 038702 (2009).
9. Salganik, M. J., Dodds, P. S. & Watts, D. J. Experimental study of inequality and unpredictability in an artificial cultural market. *Science* **311**, 854–856 (2006).
10. Aral, S., Muchnik, L. & Sundararajan, A. Engineering social contagions: optimal network seeding in the presence of homophily. *Netw. Sci.* **1**, 125–153 (2013).
11. Aral, S. & Walker, D. Identifying influential and susceptible members of social networks. *Science* **337**, 337–341 (2012).
12. Kwak, H., Lee, C., Park, H. & Moon, S. What is Twitter, a social network or a news media? In *Proc. 19th International Conference on World Wide Web* 591–600 (ACM, 2010).
13. Gruhl, D., Guha, R., Liben-Nowell, D. & Tomkins, A. Information diffusion through blogspace. In *Proc. 13th International Conference on World Wide Web* 491–501 (ACM, 2004).
14. Liben-Nowell, D. & Kleinberg, J. Tracing information flow on a global scale using Internet chain-letter data. *Proc. Natl Acad. Sci. USA* **105**, 4633–4638 (2008).
15. Goel, S., Anderson, A., Hofman, J. & Watts, D. J. The structural virality of online diffusion. *Manag. Sci.* **62**, 180–196 (2015).
16. Goel, S., Watts, D. J. & Goldstein, D. G. The structure of online diffusion networks. In *Proc. 13th ACM Conference on Electronic Commerce* 623–638 (ACM, 2012).
17. Pei, S., Muchnik, L., Tang, S., Zheng, Z. & Makse, H. A. Exploring the complex pattern of information spreading in online blog communities. *PLoS ONE* **10**, e0126894 (2015).
18. Muchnik, L., Aral, S. & Taylor, S. J. Social influence bias: a randomized experiment. *Science* **341**, 647–651 (2013).
19. Bapna, R., Ramaprasad, J., Shmueli, G. & Umyarov, A. One-way mirrors in online dating: a randomized field experiment. *Manag. Sci.* **62**, 3100–3122 (2016).
20. Eckles, D., Kizilcec, R. F. & Bakshy, E. Estimating peer effects in networks with peer encouragement designs. *Proc. Natl Acad. Sci. USA* **113**, 7316–7322 (2016).
21. Centola, D. An experimental study of homophily in the adoption of health behavior. *Science* **334**, 1269–1272 (2011).
22. Goldenberg, J., Libai, B. & Muller, E. Talk of the network: a complex systems look at the underlying process of word-of-mouth. *Mark. Lett.* **12**, 211–223 (2001).
23. Watts, D. J. A simple model of global cascades on random networks. *Proc. Natl Acad. Sci. USA* **99**, 5766–5771 (2002).
24. Castellano, C., Fortunato, S. & Loreto, V. Statistical physics of social dynamics. *Rev. Mod. Phys.* **81**, 591 (2009).
25. Zhang, Z. K. et al. Dynamics of information diffusion and its applications on complex networks. *Phys. Rep.* **651**, 1–34 (2016).
26. Pei, S. & Makse, H. A. Spreading dynamics in complex networks. *J. Stat. Mech.* **2013**, P12002 (2013).
27. Domingos, P. & Richardson, M. Mining the network value of customers. In *Proc. 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 57–66 (ACM, 2001).
28. Kempe, D., Kleinberg, J. & Tardos, É. Maximizing the spread of influence through a social network. In *Proc. 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 137–146 (ACM, 2003).
29. Kitsak, M. et al. Identification of influential spreaders in complex networks. *Nat. Phys.* **6**, 888–893 (2010).
30. Lü, L. et al. Vital nodes identification in complex networks. *Phys. Rep.* **650**, 1–63 (2016).
31. Hu, Y. et al. Local structure can identify and quantify influential global spreaders in large scale social networks. *Proc. Natl Acad. Sci. USA* **115**, 7468–7472 (2018).
32. Aral, S. & Dhillon, P. S. Social influence maximization under empirical influence models. *Nat. Hum. Behav.* **2**, 375 (2018).
33. Pei, S., Wang, J., Morone, F. & Makse, H. A. Influencer identification in dynamical complex systems. *J. Complex Netw.* **8**, cnz029 (2020).
34. Moreno, Y., Nekovee, M. & Pacheco, A. F. Dynamics of rumor spreading in complex networks. *Phys. Rev. E* **69**, 066130 (2004).
35. Kermack, W. O. & McKendrick, A. G. A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond. A.* **115**, 700–721 (1927).
36. Pastor-Satorras, R., Castellano, C., Van Mieghem, P. & Vespignani, A. Epidemic processes in complex networks. *Rev. Mod. Phys.* **87**, 925 (2015).
37. Iribarren, J. L. & Moro, E. Branching dynamics of viral information spreading. *Phys. Rev. E* **84**, 046116 (2011).
38. Centola, D. & Macy, M. Complex contagions and the weakness of long ties. *Am. J. Sociol.* **113**, 702–734 (2007).
39. De Domenico, M., Lima, A., Mougel, P. & Musolesi, M. The anatomy of a scientific rumor. *Sci. Rep.* **3**, 2980 (2013).
40. Stephen, A. T., Dover, Y., Muchnik, L. & Goldenberg, J. Pump it out! The effect of transmitter activity on content propagation in social media. *Saïd Business School WP* https://doi.org/10.2139/ssrn.2897582 (2017).
41. Rodriguez, M. G., Gummadi, K. & Schoelkopf, B. Quantifying information overload in social media and its impact on social contagions. In *Proc. 8th International AAAI Conference on Weblogs and Social Media* 170–179 (AAAI, 2014).
42. Weng, L., Flammini, A., Vespignani, A. & Menczer, F. Competition among memes in a world with limited attention. *Sci. Rep.* **2**, 335 (2012).
43. Gleeson, J. P., Ward, J. A., O'Sullivan, K. P. & Lee, W. T. Competition-induced criticality in a model of meme popularity. *Phys. Rev. Lett.* **112**, 048701 (2014).
44. Feng, L. et al. Competing for attention in social media under information overload conditions. *PLoS ONE* **10**, e0126090 (2015).
45. Hodas, N. O. & Lerman, K. How visibility and divided attention constrain social contagion. In *Proc. 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing* 249–257 (IEEE, 2012).
46. Lerman, K. Information is not a virus, and other consequences of human cognitive limits. *Future Internet* **8**, 21 (2016).
47. Muchnik, L. et al. Origins of power-law degree distribution in the heterogeneity of human activity in social networks. *Sci. Rep.* **3**, 1783 (2013).
48. Barabási, A. L. The origin of bursts and heavy tails in human dynamics. *Nature* **435**, 207–211 (2005).
49. Akbarpour, M. & Jackson, M. O. Diffusion in networks and the virtue of burstiness. *Proc. Natl Acad. Sci. USA* **115**, E6996–E7004 (2018).
50. Goldenberg, J., Han, S., Lehmann, D. R. & Hong, J. W. The role of hubs in the adoption process. *J. Mark.* **73**, 1–13 (2009).
51. Mønsted, B., Sapieżyński, P., Ferrara, E. & Lehmann, S. Evidence of complex contagion of information in social media: an experiment using Twitter bots. *PLoS ONE* **12**, e0184148 (2017).
52. Weng, L., Menczer, F. & Ahn, Y. Y. Virality prediction and community structure in social networks. *Sci. Rep.* **3**, 2522 (2013).
53. Cheng, J., Adamic, L., Dow, P. A., Kleinberg, J. M. & Leskovec, J. Can cascades be predicted? In *Proc. 23rd International Conference on World Wide Web* 925–936 (ACM, 2014).
54. Pei, S., Muchnik, L., Andrade, J. S. Jr, Zheng, Z. & Makse, H. A. Searching for superspreaders of information in real-world social media. *Sci. Rep.* **4**, 5547 (2014).
55. Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap* (CRC Press, 1994).

Acknowledgements

B.Z. is funded by Natural Science Foundation of China (grant no. 61503159) and Jiangsu University Overseas Training Programme. S.P. is supported by NIH NIGMS grant no. 5U01GM110748. L.M. is supported by Israel Science Foundation grant no. 1777/17. X.M. and H.E.S. are supported by NSF Grant PHY-1505000 and DTRA grant no. HDTRA-1-14-1-0017. X.X. is supported by National Natural Science Foundation of China (grant no. 61773091). A.S. wishes to thank the Ariel Cyber Innovation Centre in conjunction with the Israel National directorate in the Prime Minister's Office for their support. S.H. thanks the Italian Ministry of Foreign Affairs and International Cooperation jointly with the Israeli Ministry of Science, Technology, and Space (MOST); the Israel Science Foundation, ONR, the Japan Science Foundation with MOST, BSF-NSF, ARO, the Bar-Ilan University Centre for Research in Applied Cryptography and Cyber Security and DTRA (grant no. HDTRA-1-19-1-0016) for financial support.

The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

All authors designed the research. B.Z. and S.P. performed the experiments and analysis. B.Z., S.P. and L.M. curated data. B.Z., S.P. and L.M. wrote the first draft of the manuscript. X.M., X.X., A.S., S.H. and H.E.S. reviewed and edited the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-020-00945-1.

Correspondence and requests for materials should be addressed to B.Z. or S.P.

Peer review information Primary handling editor: Aisha Bradshaw.

nature research

Corresponding author(s): Sen Pei

Last updated by author(s): 07/11/2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data were collected from online depositories. LiveJournal data were stored in MySQL.
Data analysis	Data were analyzed using Python 2.7.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The Weibo and Twitter data are publicly available. The LiveJournal data are subject to restrictions due to privacy issue.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	We use information diffusion records and associated social relations in LiveJournal, Weibo and Twitter to quantitatively study the person-to-person diffusion probability.
Research sample	Retweet records and follower networks in Twitter and Weibo, URL reference records and friend relationship in LiveJournal
Sampling strategy	LiveJournal data were collected through crawling all public accessible pages. Weibo tweets were randomly sampled. Twitter data were collected using terms related to the discovery of Higgs boson.
Data collection	Weibo and Twitter data were downloaded from public repositories. LiveJournal data were collected through crawling all public accessible pages.
Timing	LiveJournal data were collected during Feb. 14th, 2010 to Nov. 21st, 2011. Time of collection for Weibo is not reported. The Twitter dataset was collected by monitoring retweets on Twitter about the discovery of Higgs boson around Jul. 7th, 2012, when the news was announced.
Data exclusions	No data were excluded.
Non-participation	Not applicable.
Randomization	Not applicable.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging