Effective spreading from multiple leaders identified by percolation in the susceptible-infected-recovered (SIR) model

2017 New J. Phys. 19 073020

(http://iopscience.iop.org/1367-2630/19/7/073020)

New Journal of Physics
The open access journal at the forefront of physics

Deutsche Physikalische Gesellschaft DPG
IOP Institute of Physics

Published in partnership with: Deutsche Physikalische Gesellschaft and the Institute of Physics

OPEN ACCESS

RECEIVED
7 December 2016

REVISED
23 May 2017

ACCEPTED FOR PUBLICATION
2 June 2017

PUBLISHED
20 July 2017

PAPER

Effective spreading from multiple leaders identified by percolation in the susceptible-infected-recovered (SIR) model

Shenggong Ji[1,2], **Linyuan Lü**[3,4], **Chi Ho Yeung**[5] **and Yanqing Hu**[1,6,7]

[1] School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510006, People's Republic of China
[2] School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, People's Republic of China
[3] Alibaba Research Center for Complexity Sciences, Hangzhou Normal University, Hangzhou 310036, People's Republic of China
[4] Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 610054, People's Republic of China
[5] Department of Science and Environmental Studies, The Education University of Hong Kong, Hong Kong, People's Republic of China
[6] Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, Guangzhou 510006, People's Republic of China
[7] Big Data Research Center, University of Electronic Science and Technology of China, Chengdu 611731, People's Republic of China

E-mail: yanqing.hu.sc@gmail.com

Keywords: complex networks, percolation, spreading, social networks

Supplementary material for this article is available online

Abstract

Social networks constitute a new platform for information propagation, but its success is crucially dependent on the choice of spreaders who initiate the spreading of information. In this paper, we remove edges in a network at random and the network segments into isolated clusters. The most important nodes in each cluster then form a set of influential spreaders, such that news propagating from them would lead to extensive coverage and minimal redundancy. The method utilizes the similarities between the segmented networks before percolation and the coverage of information propagation in each social cluster to obtain a set of distributed and coordinated spreaders. Our tests of implementing the susceptible-infected-recovered model on Facebook and Enron email networks show that this method outperforms conventional centrality-based methods in terms of spreadability and coverage redundancy. The suggested way of identifying influential spreaders thus sheds light on a new paradigm of information propagation in social networks.

1. Introduction

The development of social networks has had a great impact on our lifestyles, from making friends to dating, from working to shopping. They become more essential as we are increasing our dependence on them to gather information. Compared with search engines which are based on isolated queries, collecting information through leveraging the individual specialties in social networks leads us to useful information given by experts in disparate fields, and thus increases both the quality and the diversity of acquired information. By the same token, influential individuals can also be used to spread information. The key to success is to identify the most influential spreaders in the network. However, it is difficult to identify them as there are usually just a few individuals capable of propagating a piece of news to a large number of users [1]. For example, while socially significant users are rare in the Twitter network, their messages, and blogs can spread quickly throughout the whole network [2, 3].

Some simple methods have been proposed to identify optimal spreaders. For instance, degree centrality suggests that nodes with higher degree are more influential than others [4]. On the other hand, the location of a node in a network and the influence of its neighbors are also important. For instance, a node with a small number of highly influential neighbors located at the center of the network may be more influential than a node having a larger number of less influential neighbors. Kitsak *et al* [5] thus proposed a coarse-grained method to use the *k*-core decomposition to quantify the influence of a node, based on the assumption that news initiated at

nodes in higher shells is likely to spread more extensively. However, this location-based method is invalid for tree-like networks where all nodes are in the same shell. Recently, Morone and Makse mapped the influencer identification problem onto optimal percolation and proposed a metric called 'collective influence' to find the solution [6] . Their method can find a class of strategic influencers which outrank the hubs in the networks. Some distance-based global metrics such as betweenness [7] and closeness [8] are also suggested which can lead to extensive propagation, but due to high computational complexity, they are not practical for large-scale social networks. Other centralities such as LocalRank were also suggested [9]. We remark that although the selected spreaders by different methods may be influential in specific spreading models, these results are usually sensitive to the chosen spreading mechanism [10, 11]. In this paper, we only study the susceptible-infected-recovered (SIR) models, which well describe some aspects of the information spreading process in social media [12–15].

Simple but sub-optimal protocols have been applied to social media such as QQ, BBS, and Blog to find the key spreaders who can trigger the 'tipping point' in social marketing to promote commercial products. Specifically, if one can convince a set of influential users to adopt a new product, one may induce a large cascade of purchases, as these initial buyers propagate their compliments of the product along the network. Unlike the forementioned methods which identify a set of independent spreaders according to their centralities, our goal is to find a coordinated set of individuals such that their combined impact is greatest, leading to much more extensive propagation of information. However, identifying the optimal combination of spreaders is indeed a difficult task, both conceptually and computationally [16].

In this paper, we utilize the similarities [15, 17] between bond percolation and information propagation to identify a group of influential spreaders. By removing edges at random until percolation ceases, individual isolated clusters are formed. Due to the correspondence between percolation and information transmission, the emergence of such clusters implies that news can be effectively propagated within the clusters but not across the clusters. Initiating a piece of news on the most influential user identified by degree centrality in each cluster is thus an effective way to distribute the news within the cluster. Since such a process is static and requires much less computation power than the dynamical spreading of news, a lot of segmented states can be generated and averaged to give a more accurate result on the segmentation of social clusters as well as their corresponding influential spreaders.

By testing our method on Facebook and Enron email networks, we show that in addition to a higher computational efficiency, our method outperforms other simple heuristics based on local and global centrality in terms of propagation coverage and coverage redundancy of the selected spreaders. This is consistent with the old saying that the power of a typical group exceeds that of a single most competent individual. Moreover, we find that the average degree of the users selected by our method is lower, which implies a lower cost in identifying the spreaders when compared to other methods. We also identify the different characteristics of spreaders who are most effective in promoting niche or popular items in order to maximize coverage. All these results lead to insights into the design of viral marketing strategies and a new paradigm for information propagation.

2. The model

Spreading dynamics with the involvement of humans can be mainly classified into two classes: one is the spreading of infectious diseases which requires physical contact, and the other is the spreading of information, including opinions and rumors where physical contact is not required [18]. Due to the similarity between epidemic and information spreading, well-established models of epidemic spreading are widely used to describe the propagation of information [12–15, 19, 20].

In particular, the susceptible-infected-recovered (SIR) model is one representative [12–15]. Individuals in this model are classified into three states: susceptible (S, does not carry the disease and will not infect others but can be infected), infected (I, carries the disease and can infect others), recovered (R, either dead or recovered from the disease and immune to further infection). The simulation runs in discrete time steps. At each time step, an infective individual transmits the disease to his or her neighbors with probability β and will recover with probability γ. Then the SIR transmissibility is $p = \beta/\gamma$. The process stops when there is no infected node anymore. When applying the SIR model to mimic information spreading, a susceptible person (S) in the model is analogous to an individual who is not aware of the information. An infected person (I) is analogous to an individual who is aware of the information and will pass it to his/her neighbors. A recovered person (R) is analogous to an individual who loses his/her interest and will never pass the information on again.

Newman [15] studied in detail the relationship between the static properties of the SIR model and bond percolation phenomena on networks, and remarked that the SIR model with transmissibility p is equivalent to a bond percolation model with bond occupation probability p on the network. After removing the other edges, a number of clusters are formed. It is clear that the ultimate size of the SIR epidemic outbreak triggered by a single initially infected node is precisely the size of the cluster that the initial node belongs to. Apparently, the nodes in

Table 1. The basic characteristics of Facebook and Enron email networks. We denote by $|V|$ and $|E|$ the number of nodes and edges, respectively, C the clustering coefficient [26] and r the assortative coefficient [27]. We denote by $\langle k\rangle$ the average degree, $\langle d\rangle$ the average shortest distance, and H the degree heterogeneity, such that $H = \langle k^2\rangle/\langle k\rangle^2$.

Networks	$\lvert V\rvert$	$\lvert E\rvert$	C	r	$\langle k\rangle$	$\langle d\rangle$	H
Email	33696	180811	0.510	−0.117	10.732	4.025	13.266
Facebook	59691	1456818	0.228	0.175	24.4060	4.335	3.348

the same cluster are expected to have the same coverage. A review article on epidemic processes in complex networks can be found in [17].

Our method is then devised in relation to the bond percolation model [15, 21–24] as follows. Given an undirected network $G(V, E)$ where V represents the set of nodes (i.e. users in social networks) and E represents the set of edges (i.e. connection in terms of communication, friendship, or other kinds of interactions), all edges are first removed and each individual edge is then recovered with a probability p. All links are removed when $p = 0$, and when p increases, more links are recovered and clusters start to form and merge with each other. For a network containing N nodes, a giant component of size $O(N)$ emerges only when p is larger than a critical threshold $p = p_c$, and this phenomenon is called percolation. In this paper, we will call those states with isolated clusters the *segmented states*. In the context of information propagation, since an edge between two nodes appears with a probability p, the value p can be considered as the transmissibility of information from one node to another.

To find the most influential group, we identify the W most influential spreaders in the network by utilizing the segmented states where some of the edges are removed. Assume that there are m isolated clusters in a segmented state after one realization of link recovery, and denote by S_i the size of cluster i, for $i = 1, 2, \cdots, m$. We introduce a tunable parameter L, which is usually equal to or larger than W. If $L \leqslant m$, we choose the top-L largest clusters and assign one unit of 'leader score' to the largest degree node in each cluster. If there are many nodes with the largest degree, we randomly assign the score to one of them. If $m < L \leqslant 2m$, we first choose the highest degree node in each of the top-m largest clusters, and the rest of the $L - m$ nodes are chosen to be those with the second largest degree respectively from the top-$(L - m)$ largest clusters. If $L > 2m$, we will choose the next largest degree nodes in each cluster following the same selection rules. After M times of different trials of link recovery, all nodes are ranked according to their scores in a descending order and those W nodes with the highest leader scores are suggested to be the set of initial spreaders. For the sake of simplicity, we set $L = W$ and have tested and found that the results are not sensitive to L. The dependence of the results on L is shown in supplementary figure S1 available online at stacks.iop.org/NJP/19/073020/mmedia.

In other words, our suggested method draws an analogy with percolation to identify individual social clusters in the network where news can be effectively propagated within clusters but not across clusters. These isolated clusters in the segmented state thus have a direct correspondence to propagation coverage when one spreads news from an initial spreader in each of the clusters. Unlike most other methods which usually identify a group of influential spreaders that are not evenly distributed to cover the whole network, our procedures segment the network into non-overlapping components such that the identified spreaders are well distributed, enjoying reduced redundancy when compared to a set of uncoordinated spreaders. These differences make our method unique compared to other methods.

3. Results

3.1. Datasets

We consider two social networks, namely the Facebook network and the Enron email network. Their statistical features are shown in table 1.

(i) Facebook: the friendship relations in the New Orleans Facebook social network. It is a directed network, consisting of nodes which correspond to Facebook users. Each directed edge represents a post or a comment, connecting the users who are writing the comment with the users whose wall the post is posted. Since users may write multiple comments on the same wall, the network allows multiple edges connecting a node pair. Since users may also comment on their own wall, the network contains self-loops. In our experiment, we consider the network as a typical undirected network by deleting self-loops and merging multiple links into a single link. We assume that two nodes are connected if there is at least one directed link between them. The data can be freely downloaded at http://levich.engr.ccny.cuny.edu/~hmakse/soft_data.html.

(ii) Enron email network [25]: Enron's email communication network covers roughly 0.5 million email communications between a group of users. This dataset was originally open to the public, and was posted on the

internet by the Federal Energy Regulatory Commission during its investigation. Nodes of the network correspond to email addresses in the system, and if an address i sent at least one email to address j, there is an undirected edge between i and j. Note that non-Enron email addresses are considered as sources and sinks in the network, as we only observe their communications with the Enron email addresses, but not the communications between them. The data can be freely downloaded at http://snap.stanford.edu/data/email-Enron.html.

3.2. Spreadability and coverage redundancy

To quantify the performance of our method, we examine the spreadability, i.e. the propagation coverage of a piece of news from a set of W selected spreaders, by our method as well as other methods. We will use the SIR model to mimic the spreading of news, and the spreadability is defined as the ratio of recovered nodes, i.e. the size of outbreak, or the number of users who received the information to the total number of users. We remark that the transmissibility p adopted in the SIR model is the same as the probability p used to recover edges to identify the clusters in the segmented states. As a result, for a single spreader, the maximum size of the SIR outbreak triggered by this spreader is precisely the size of the cluster that it belongs to. Likewise, the maximum size of the SIR outbreak triggered by a group of spreaders in distinct clusters is the sum of the size of the clusters that these nodes belong to. For example, if we measure the maximum coverage of three selected nodes on the network with N nodes, and if the first two nodes belong to the cluster S_i, and the third one belongs to the cluster S_j, the maximum coverage of each individual node 1, 2, and 3 is respectively S_i/N, S_i/N and S_j/N, while for the whole group, the maximum coverage is $S_i + S_j$, and the spreadability is $(S_i + S_j)/N$.

We first apply our method on the Facebook network with 59691 nodes. Figure 1(a) shows the coverage obtained from 4000 initial spreaders chosen by our percolation method, compared with a set of 4000 spreaders identified by four other methods, namely the k-shell decomposition, the betweenness centrality, the collective influence (CI) method [6], and the high degree adaptive method (HDA) where the degrees of nodes are recalculated according to the updated network (see the appendix for the definition of each of these methods). The percolation method yields the highest spreadability for an arbitrary transmissibility p as shown in figure 1(a), while comparisons with other centrality measures can be found in supplementary figure S2. We show in section 4 of the supplementary data that our method also outperforms the other methods on weighted networks in terms of spreadability.

Figure 1(b) shows the degree distribution of the 4000 spreaders identified by the percolation method. When $p < p_c \approx 0.01$, the percolation method yields isolated clusters of similar size [21], and since the set of selected spreaders comes from different clusters, a wide range of degree is found among the spreaders as shown in figure 1(b). In this case, the percolation method is more likely to choose high-degree nodes, see supplementary figure S3(b), where the red stars represent the degree distribution of the 4000 selected nodes when $p = 0.008$. When $p > p_c$, the distribution becomes narrower as p increases, see the blue squares in supplementary figure S3(b). In this case, the percolation method prefers low-degree spreaders. The average original degree (i.e. the degree in the original network before edge removal) of the 4000 selected spreaders by the percolation method when $p < p_c$ is higher than that of the nodes selected when $p > p_c$. This implies that if we promote and advertise a new niche product which is regarded as difficult to accept, one can draw an analogy with the case of small transmissibility p where high-degree initial spreaders are preferred. On the other hand, for popular items which are easy to accept, one can draw an analogy with the case of large p and low-degree initial spreaders are preferred.

We then examine the cost of initializing the spreading from the selected spreaders. Information propagation sometimes induces a cost, for instance, one may need to pay the star bloggers for posting and passing on an advertisement. We assume that the direct influence of a user is equal to the number of its nearest neighbors, i.e. the degree of the user, and the difficulty of finding a user with degree k is proportional to $1/p(k)$, which can be considered the scarcity of the user. Here $p(k)$ is the occurring frequency of nodes with degree equal to k. The cost to initialize (or hire) a spreader i is proportional to his/her impact as well as scarcity, and hence the cost is assumed to be $k_i/p(k_i)$. Obviously, if $p(k_i)$ follows the power-law distribution, namely $p(k_i) \sim k^{-\alpha}$, then cost is proportional to $k^{1+\alpha}$. The larger the degree of a user, the higher the cost to hire him/her. Figure 1(c) shows the dependence of the average cost of 4000 spreaders under the parameter p, i.e. $\frac{1}{W}\sum_{i=1}^{W}\frac{k_i}{p(k_i)}$. The cost decreases abruptly at the critical point p_c, suggesting a phenomenon resembling a phase transition. It implies that when p increases just beyond p_c, the cost can be reduced substantially. We have tested that if the cost is defined as $k^{1+\alpha}$, similar results are obtained.

Besides spreadability and cost, we also examine redundancy in coverage which quantifies the efficiency of propagation. Specifically, the redundancy of a node i is defined as the number of initial spreaders who have the potential to infect node i. A method is inefficient if the initial chosen spreaders pass the same information to the same group more than once. Averaging redundancy over all infected nodes, we obtain the redundancy of the set of initial spreaders. Figure 1(e) compares the spreading redundancy of our method with the other four methods (comparisons with other centrality measures can be found in supplementary figure S4). The highest redundancy is found in the methods of k-shell and degree centrality, followed by CI, and then by betweenness centrality. Our

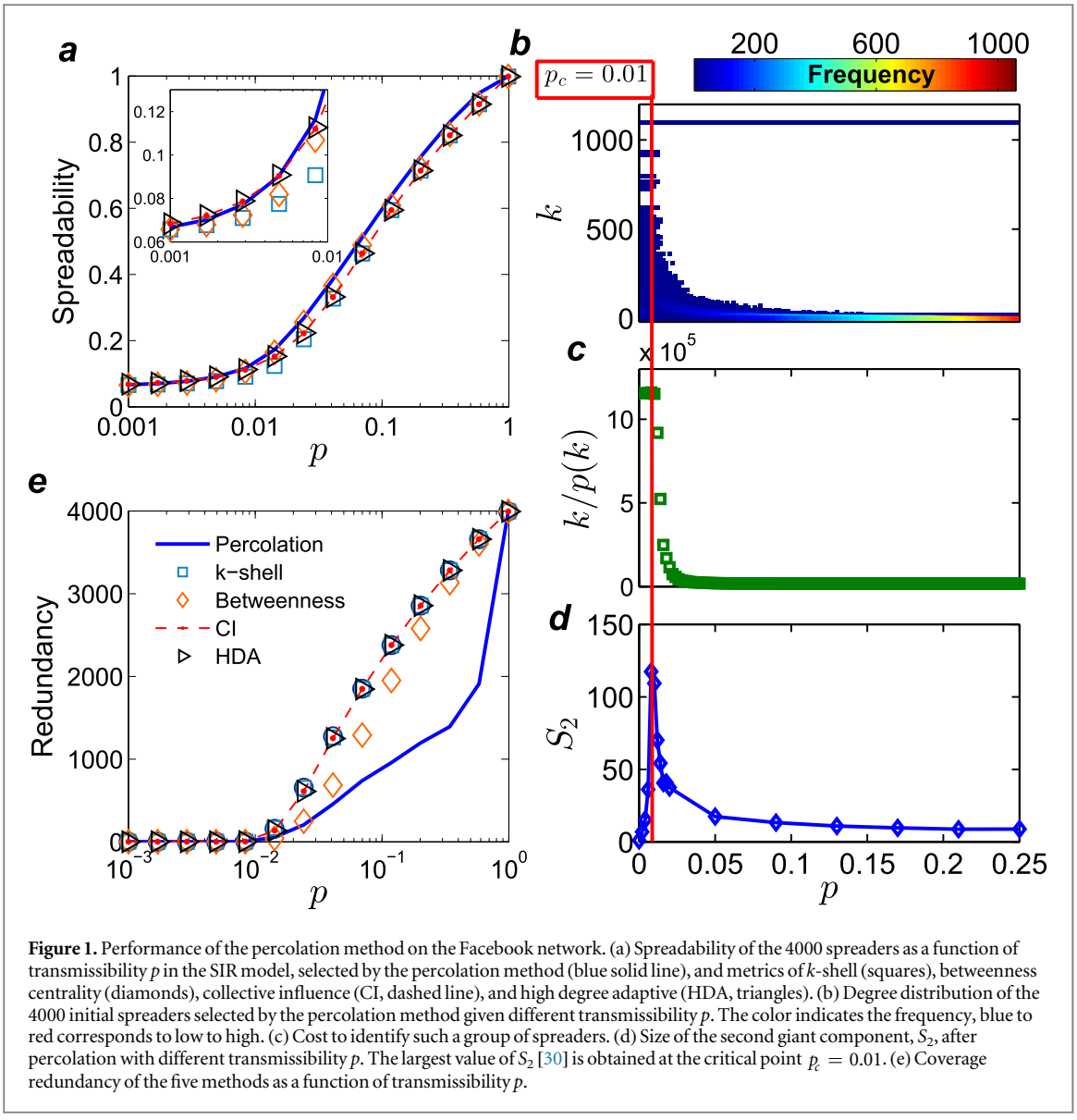


Figure 1. Performance of the percolation method on the Facebook network. (a) Spreadability of the 4000 spreaders as a function of transmissibility p in the SIR model, selected by the percolation method (blue solid line), and metrics of k-shell (squares), betweenness centrality (diamonds), collective influence (CI, dashed line), and high degree adaptive (HDA, triangles). (b) Degree distribution of the 4000 initial spreaders selected by the percolation method given different transmissibility p. The color indicates the frequency, blue to red corresponds to low to high. (c) Cost to identify such a group of spreaders. (d) Size of the second giant component, S_2, after percolation with different transmissibility p. The largest value of S_2 [30] is obtained at the critical point $p_c = 0.01$. (e) Coverage redundancy of the five methods as a function of transmissibility p.

percolation method has the lowest redundancy compared to the other four methods, since the spreaders identified by our method are usually located in different regions of the network. We also checked the Enron email network and results similar to the Facebook network are obtained, see figure 2. We also show in section 4 of the supplementary data that our method outperforms the other methods on weighted networks to reduce the redundancy in coverage.

To further examine redundancy in coverage, we applied the five methods to identify four initial spreaders on a simulated network with clear communities. There are three steps to generate a network with community structures. In our experiment, we consider a network with 2000 nodes which has four communities, each of which contains 500 nodes. First, we generate a random network of size 500 and with node degrees distributed in a power-law with exponent 2.2 using the configuration model [28]. The minimum degree is 1 and the maximum degree is $\sqrt{500} \approx 23$ [29]. Second, we repeat the above procedures to generate independently the other three networks. Finally, for each pair of sub-networks we randomly select a fraction of node pairs to connect them. As shown in table 2, the four spreaders identified by the percolation method are likely to be found in different communities. For the other methods, there are high probabilities that at least two initial spreaders are in the same community. This result is easy to understand as our method relies on the segmentation of networks into isolated clusters. In this case, the network separates into four communities and thus one spreader is found in each community.

Although the leader score is aggregated after M realizations, a different set of leaders may be generated if another set of M realizations of percolation is generated, since link recovery in our percolation-based method is stochastic in nature and the highest degree nodes in small clusters are not unique. While the other methods always suggest the same set of initial spreaders, different spreaders can be generated by reiterating our

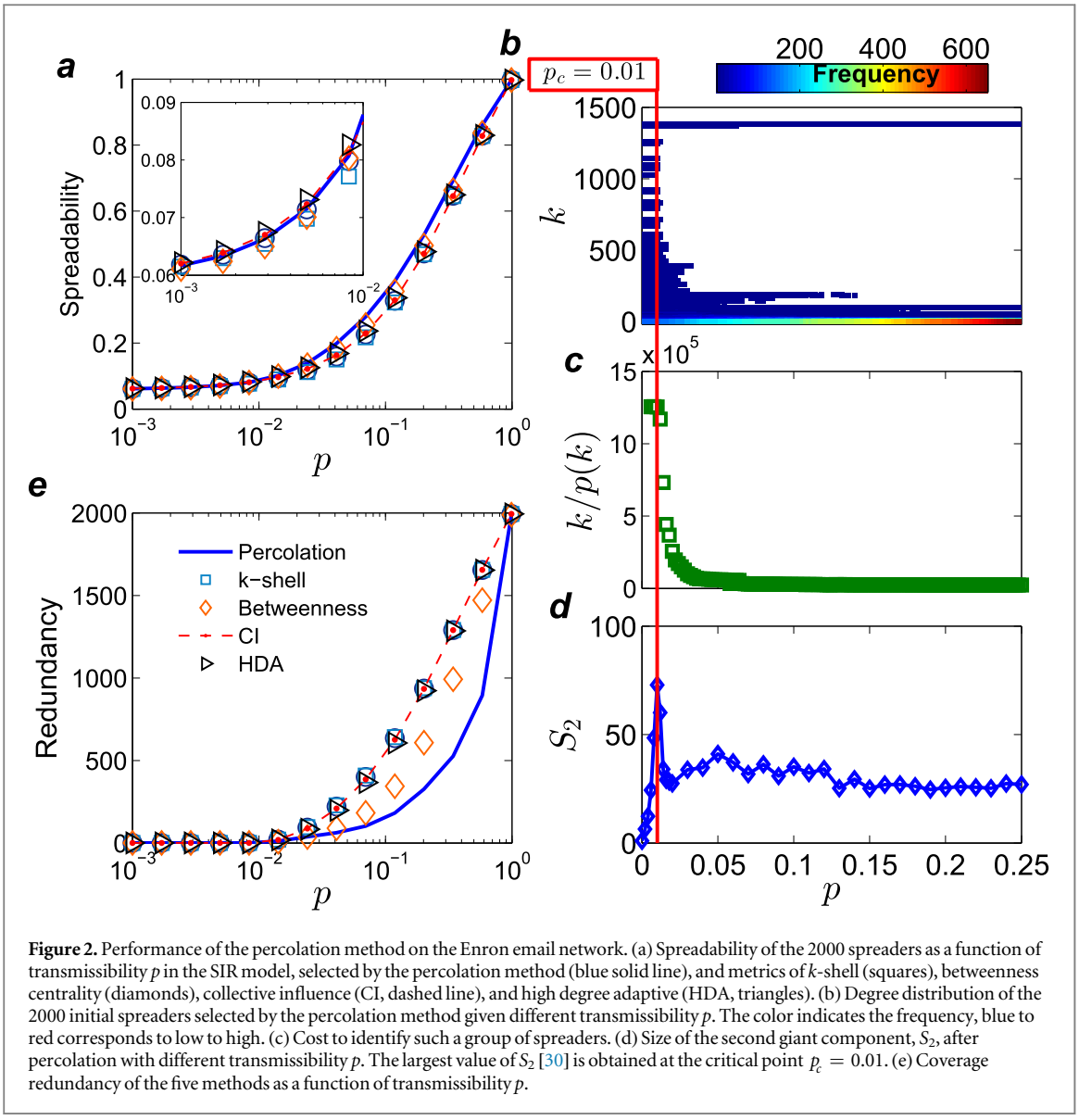
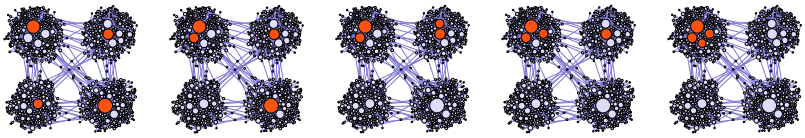


Figure 2. Performance of the percolation method on the Enron email network. (a) Spreadability of the 2000 spreaders as a function of transmissibility p in the SIR model, selected by the percolation method (blue solid line), and metrics of k-shell (squares), betweenness centrality (diamonds), collective influence (CI, dashed line), and high degree adaptive (HDA, triangles). (b) Degree distribution of the 2000 initial spreaders selected by the percolation method given different transmissibility p. The color indicates the frequency, blue to red corresponds to low to high. (c) Cost to identify such a group of spreaders. (d) Size of the second giant component, S_2, after percolation with different transmissibility p. The largest value of S_2 [30] is obtained at the critical point $p_c = 0.01$. (e) Coverage redundancy of the five methods as a function of transmissibility p.

Table 2. Percentage of different distribution of the four spreaders in a model network with four communities. For instance, the four spreaders identified by our method are found in four different communities in 93.9% of the 1000 realizations. In the percolation method, we have set $p = 0.28$ which is slightly higher than the critical point of the network $p_c \approx 0.2687 \pm 0.0152$.

Distribution of the four selected spreaders (red nodes)					
Percolation	**93.9**	6.1	0	0	0
K-shell	9.4	**43.6**	11.9	16.2	18.9
Betweenness	13.1	**64.5**	11.7	10.6	0.1
CI	10.3	**61.5**	12.2	15.2	0.8
HDA	11.7	**60.7**	13	14.2	0.4

percolation procedures, especially for large values of p. Figures 3(a) and (b) show the average number of common nodes in two different solutions of the percolation method, i.e. n_c, on Facebook and Enron email networks, respectively. It is clear that when p increases, the number of common spreaders decreases, indicating that the solutions become more diverse. This result has practical significance; in cases when some initial spreaders are offline, we can use the next best candidates as back-up spreaders without extensively losing

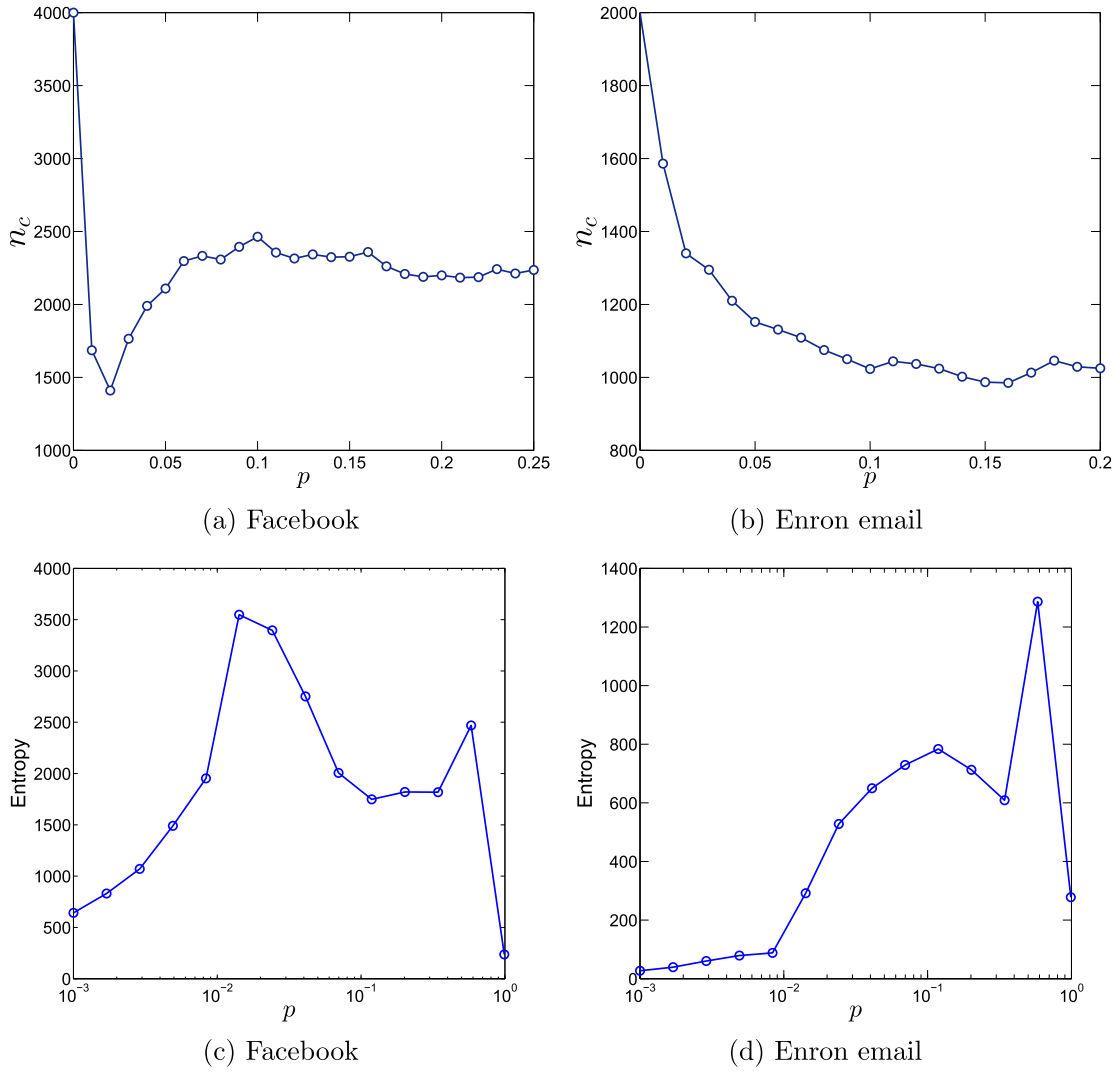


Figure 3. Diversity of the spreaders identified by the percolation method. (a), (b) Average number of common nodes in two solutions, i.e. n_c. (c), (d) Shannon entropy of the solution space. We have set $W = L = 4000$ for the Facebook network, and $W = L = 2000$ for the Enron email network.

spreadability. Compared with the other four methods, the percolation method provides higher flexibility in the choice of spreaders. We further calculate the Shannon entropy of the obtained solutions, which is defined as

$$\text{Entropy} = -\sum_{i}^{N} q_i \log q_i, \tag{1}$$

where q_i is the percentage of realizations where node i is found. Figures 3(c) and (d) show the dependence of Shannon entropy on parameter p. A non-zero Shannon entropy also indicates that various solutions are found and different sets of nodes are identified as the initial spreaders. A more non-trivial trend of the Shannon entropy is observed compared to the entropy computed by merely the number of different solutions. For instance, a peak of Shannon entropy at the intermediate values of $p > p_c$ is observed. In this case, a giant component exists in the network together with many small clusters. Depending on the random recovery of links, the set of smallest clusters is different for different solutions. When the total number of clusters is roughly equal to the number of identified spreaders, one spreader is identified for each cluster, including the smallest clusters. The different sets of smallest clusters then contribute to the different sets of identified spreaders, and hence a peak in the Shannon entropy at intermediate values of p when the number of clusters is roughly equal to L.

To further examine the difference between our method and the other methods, figure 4 shows the fraction of common spreaders with the percolation method, i.e. f. Comparison with other methods can be found in supplementary figure S5. The overlap between the percolation method and the degree centrality method reaches the highest value at the critical point $p_c = 0.01$ and then sharply decreases to less than 5% when $p = 0.03$. This is because when p increases, most high-degree nodes are replaced by nodes with smaller degree, and there are

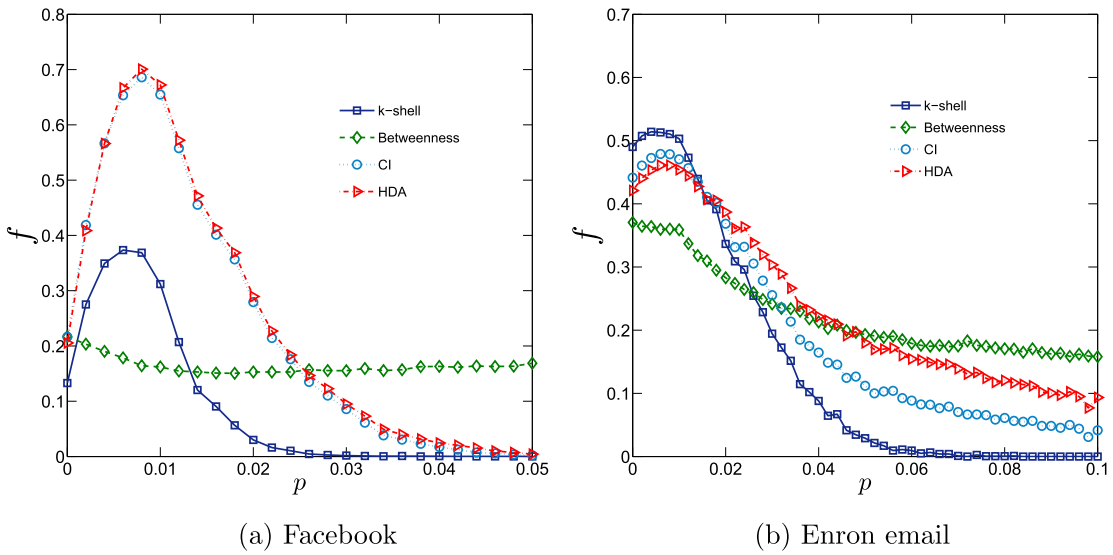


Figure 4. Fraction of common spreaders identified by the percolation method and the other three methods on the Facebook network. We have set $W = L = 4000$ for the Facebook network, and $W = L = 2000$ for the Enron email network.

many sets of spreaders generated from different realizations of the percolation method as we have discussed in figure 3.

We show in figures 5 and 6 the spreadability-cost profile of the group of top-W selected spreaders. Similar results are found for the two networks. Four cases are presented, namely $p = 0.008 < p_c$, $p = 0.01 = p_c$, $p = 0.012$, and $p = 0.02 > p_c$. As we can see, the percolation method is most cost effective in a large range of spreader cost, i.e. with the highest spreadability given a selected spreader of the same cost. On the other hand, betweenness identified low-cost spreaders with high spreadability, but their spreadability is limited to a small maximum value compared to the other periods. In this case, the spreader with maximum spreadability is always identified by the percolation method. Besides the two real networks, we also investigated scale-free networks. Similar results are found; see supplementary figures S6–S11.

4. Discussion

As we can see, social networks constitute a new platform to propagate information. Unlike the usual practice where the networks are used by uncoordinated individuals to share their own message, controlled spreading of information can be implemented via the networks. To quantify its performance, one can measure the coverage, the redundancy in propagation, and the cost in identifying appropriate initial spreaders. Yet these measures of performance are largely dependent on the choice of users who start the propagation, and there is not a single protocol which achieves optimality in all these dimensions. These difficulties of identifying influential spreaders make controlled information propagation via social networks solely theoretical.

To tackle the challenge, we draw an analogy between percolation processes and information propagation to develop a method which gives rise to a low-cost, minimally redundant set of initial spreaders capable of achieving large propagation coverage. Our method was tested on Facebook and Enron email networks, where favorable results over centrality-based methods were obtained. When compared to uncoordinated spreaders identified by these conventional methods, the spreaders identified by our method are evenly distributed within the network which greatly increases the propagation coverage and reduces its redundancy. Such coordination of spreaders is essential and can only be obtained using the suggested percolation procedures.

The success of this method is not just a coincidence, since it utilizes similarities between percolation and information propagation. By removing edges at random until percolation ceases, we identify individual isolated clusters where news can be effectively propagated within the clusters but not across the clusters. Specific spreaders at the center of these clusters are then identified to be the influential initial spreaders in the original network. By initiating news propagating from this set of spreaders, coverage is increased and redundancy is reduced compared to conventional centrality methods. Percolation is thus at the center of our method instead of a mere analogy.

The remaining question is practicality. As we have shown in the appendix, the computational complexity of our method is $O(M|V|)$ with M the number of realizations, which is a favorable characteristic for applying on real systems as its complexity scales linearly with system size. Once the set of important initial spreaders is

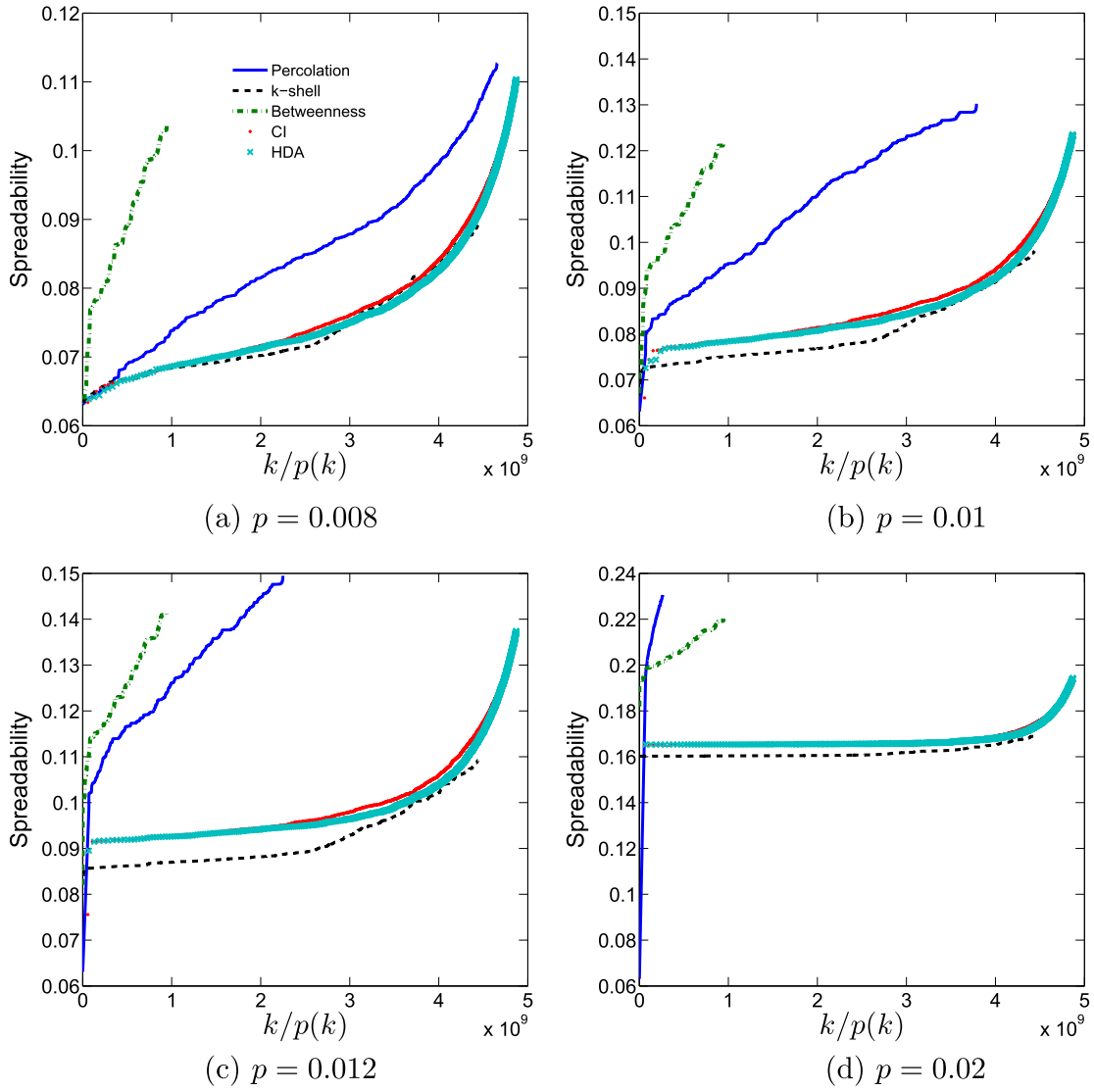


Figure 5. Spreadability as a function of the cost of the group of top-W spreaders identified by the various methods on the Facebook network. We selected 4000 nodes and set $L = 4000$ in the percolation method. Here we present four typical examples. For a given p, we select 4000 top-ranked nodes according to the different methods. For example, to obtain the results of the percolation method (i.e. blue solid line), we plot the total cost of the top-W ($W = 1, 2, 3, \cdots, 4000$) nodes as x-axis values and their corresponding spreadability (as a group) as y-axis values. k refers to the degree of a spreader and $p(k)$ denotes the probability of a spreader with degree k in the entire social network.

identified, a coordinator just has to connect to these users and pass them the news; information will propagate quickly throughout the network. Of course, a lot of details and practical issues are omitted in this simple description, but our results shed light on a completely new paradigm of information propagation. Further research along this line may revolutionize our way of spreading and gathering information in the near future.

Acknowledgments

This work is partially supported by NSFC Grant Nos. 61203156, 11622538, 61673150, and the Fundamental Research Funds for the Central Universities Grant No. 2682014RC17. LL acknowledges Zhejiang Provincial Natural Science Foundation of China under Grant LR16A05000. CHY acknowledges the Research Grant Council of Hong Kong (grant number ECS 28300215 and GRF 18304316). YH is supported by the Hundred-Talent Program of the Sun Yat-sen University, and the Chinese Fundamental Research Funds for the Central Universities Grant 16lgjc84. SJ and LL contributed equally to this work.

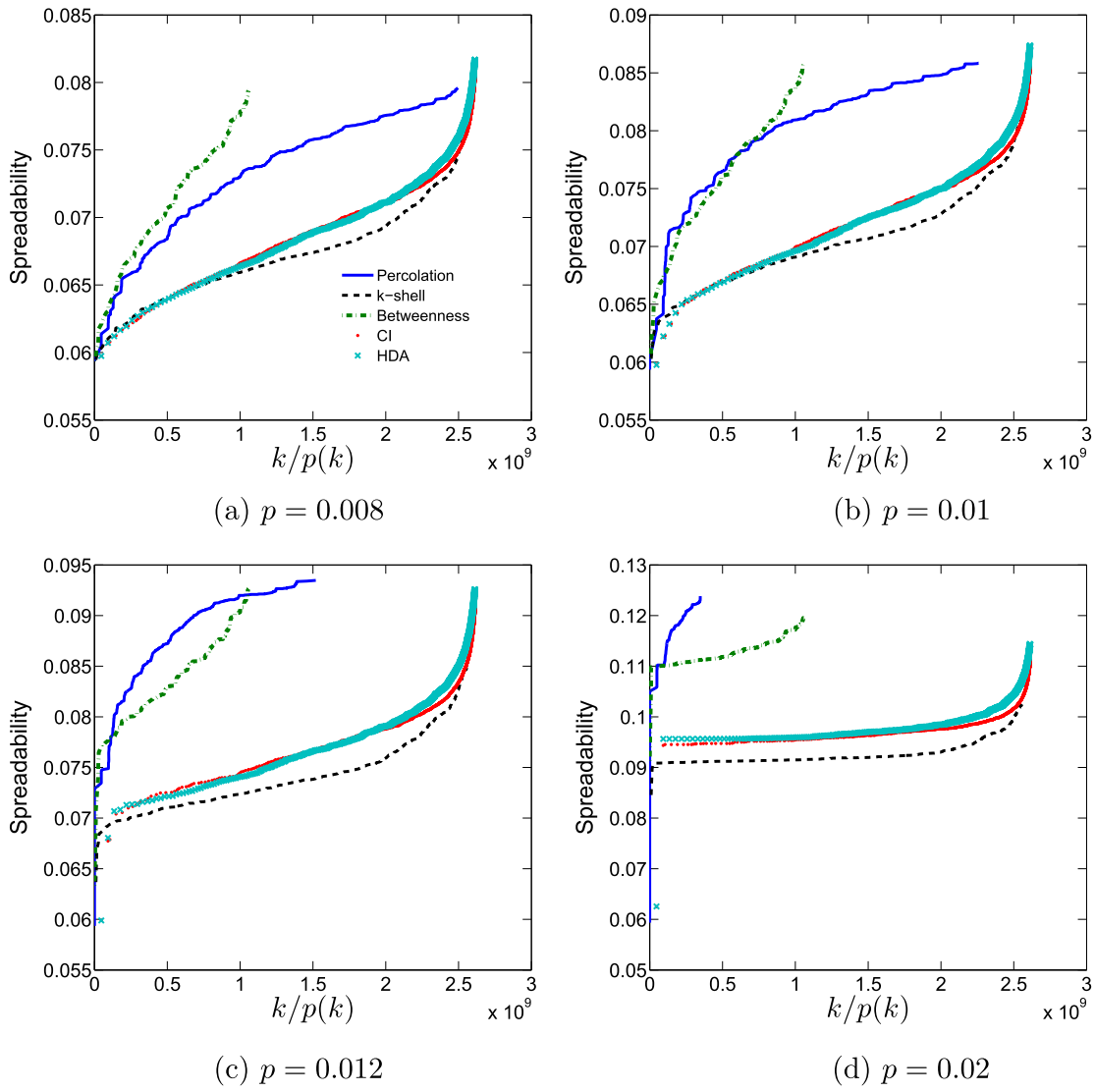


Figure 6. Spreadability as a function of the cost of the group of top-W spreaders identified by the various methods on the Enron email network. We selected 2000 nodes and set $L = 2000$ in the percolation method. Here we present four typical examples. For a given p, we select 2000 top-ranked nodes according to the different methods. For example, to obtain the results of the percolation method (i.e. blue solid line), we plot the total cost of the top-W ($W = 1, 2, 3, \cdots, 2000$) nodes as x-axis values and their corresponding spreadability (as a group) as y-axis values. k refers to the degree of a spreader and $p(k)$ denotes the probability of a spreader with degree k in the entire social network.

Appendix

A.1. Methods for comparison

To identify the most influential spreaders, various centrality measures have been proposed. The simplest method is degree centrality, which we compare our result with. Degree centrality is a straightforward and efficient metric. It assumes that a node with more nearest neighbors has a higher influence. However, node degree can only reflect its direct influence and not the indirect influence triggered by its nearest neighbors. For example, a node of small degree, but with a few highly influential neighbors may be more influential than a node having a larger number of less influential neighbors. In this paper, we employed the adaptive version of degree centrality as one of the baseline methods, namely the high degree adaptive (HDA) approach which recalculates node degree after the removal of links in the network. We compare the high degree (HD) method with the high degree adaptive (HDA) method and find that the adaptive method performs slightly better than the static high degree strategy (see supplementary figure S2).

The second method we used for comparison is k-shell decomposition. Recent research shows that the location of a node in a network may play a more important role than its degree. A node located at the center of the network is more influential than a node having a larger number of less influential neighbors. Similar to this rationale, Kitsak *et al* [5] proposed a coarse-grained method by using the method of k-core decomposition to

quantify the influence of a node based on the assumption that nodes in the same shell have similar influence, and nodes in higher-level shells are likely to infect more nodes.

The third method is betweenness which is one of the most popular geodesic-path-based ranking measures. It is defined as the fraction of shortest paths between all node pairs that pass through the node of interest. Betweenness is, in some sense, a measure of the influence of a node in terms of its role in spreading information [31, 32]. For a network $G = (V, E)$ with $n = |V|$ nodes and $m = |E|$ edges, the betweenness centrality of node v, denoted by $B(v)$ is [7, 33]

$$B(v) = \Sigma_{s\neq v, s\neq t, v\neq t} \frac{g_{st}(v)}{g_{st}} \tag{2}$$

where g_{st} is the number of shortest paths between nodes s and t, and $g_{st}(v)$ denotes the number of shortest paths between nodes s and t which pass through node v.

The last method we compare our method with is the 'collective influence' (CI) method proposed by Morone and Makse [6]. Define Ball(i, l) as the set of nodes inside a ball of radius l (defined as the shortest path) around node i, ∂Ball(i, l) is the frontier of the ball. Then the CI index of node i at level l is defined as

$$CI_l(i) = (k_i - 1) \sum_{j\in\partial \mathrm{Ball}(i,l)} (k_j - 1), \tag{3}$$

where k_i is the degree of node i. Here we set $l = 3$.

A.2. Computational complexity

Given a network $G(V, E)$, there are four steps to find the W influential spreaders by the percolation method. Firstly, all the edges are first removed and then recovered with a probability p; we then obtain a new network G' in segmented state. The required computational complexity is $O(|E|)$. Secondly, we find the strongly connected components of G' using Tarjan's algorithm [34] which has a complexity of $O(|V| + |E|)$. Thirdly, we select one node with the highest degree in each of the L largest components and assign one score to the selected nodes. This complexity for the procedures is $O(L \times |V|)$. Repeating the above three steps for different realizations, we rank the nodes according to their scores in descending order, and the top-W nodes are chosen to be the most influential spreaders. The different realizations of the percolation process can be computed in parallel and the complexity of each implementation is $O(|E| + |V| + |E| + L \cdot |V|)$. Considering $\langle k\rangle = \frac{2|E|}{|V|}$, then the complexity is $O[(\langle k\rangle + L + 1) \cdot |V|]$. Since $\langle k\rangle \ll |V|$ in real networks, then we have $O[(\langle k\rangle + L + 1) \cdot |V|] \sim O(|V|)$, i.e. the complexity of one realization of our method grows linearly with system size. Assume there are M realizations, the total complexity would be $O(M|V|)$.

References

[1] Albert R, Jeong H and Barabási A-L 2000 Error and attack tolerance of complex networks *Nature* **406** 378–82
[2] Weng J, Lim E-P, Jiang J and He Q 2010 Twitterrank: finding topic-sensitive influential twitterers *Proc. 3rd ACM Int. Conf. on Web Search and Data Mining* 261–70
[3] Hu Y, Havlin S and Makse H A 2014 Conditions for viral influence spreading through multiplex correlated social networks *Phys. Rev.* X **4** 021031
[4] Wasserman S 1994 *Social Network Analysis: Methods and Applications* vol 8 (Cambridge: Cambridge University Press)
[5] Kitsak M, Gallos L K, Havlin S, Liljeros F, Muchnik L, Stanley H E and Makse H A 2010 Identification of influential spreaders in complex networks *Nat. Phys.* **6** 888–93
[6] Morone F and Makse H A 2015 Influence maximization in complex networks through optimal percolation *Nature* **524** 65
[7] Freeman L C 1977 A set of measures of centrality based on betweenness *Sociometry* **40** 35–41
[8] Sabidussi G 1966 The centrality index of a graph *Psychometrika* **31** 581–603
[9] Chen D, Lü L, Shang M-S, Zhang Y-C and Zhou T 2012 Identifying influential nodes in complex networks *Physica* A **391** 1777–87
[10] Singh P, Sreenivasan S, Szymanski B K and Korniss G 2013 Threshold-limited spreading in social networks with multiple initiators *Sci. Rep.* **3** 2330
[11] Pei S, Teng X, Shaman J, Morone F and Makse H A 2016 Efficient collective influence maximization in threshold models of behavior cascading with first-order transitions *Sci. Rep.* **7** 45240
[12] Daley D J and Kendall D G 1964 Epidemics and rumours *Nature* **204** 1118–1118
[13] Grabowski A, Kruszewska N and Kosiński R A 2008 Dynamic phenomena and human activity in an artificial society *Phys. Rev.* E **78** 066110
[14] Iribarren J L and Moro E 2011 Branching dynamics of viral information spreading *Phys. Rev.* E **84** 046116
[15] Newman M E J 2002 Spread of epidemic disease on networks *Phys. Rev.* E **66** 016128
[16] Kempe D, Kleinberg J and Tardos É 2003 Maximizing the spread of influence through a social network *Proc. 9th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* 137–46
[17] Pastor-Satorras R, Castellano C, Van Mieghem P and Vespignani A 2015 Epidemic processes in complex networks *Rev. Mod. Phys.* **87** 925
[18] Lü L, Chen D-B and Zhou T 2011 The small world yields the most effective information spreading *New J. Phys.* **13** 123005
[19] Sudbury A 1985 The proportion of the population never hearing a rumour *J. Appl. Probab.* **22** 443–6
[20] Liu Z, Lai Y-C and Ye N 2003 Propagation and immunization of infection on general networks with both homogeneous and heterogeneous components *Phys. Rev.* E **67** 031911

[21] Newman M E J, Strogatz S H and Watts D J 2001 Random graphs with arbitrary degree distributions and their applications *Phys. Rev.* E **64** 026118
[22] Feng L, Monterola C P and Hu Y 2015 The simplified self-consistent probabilities method for percolation and its application to interdependent networks *New J. Phys.* **17** 063025
[23] Yuan X, Hu Y, Stanley H E and Havlin S 2017 Eradicating catastrophic collapse in interdependent networks via reinforced nodes *Proc. Natl Acad. Sci.* **114** 3311–5
[24] Reis S D S, Hu Y, Babino A, Andrade J S Jr, Canals S, Sigman M and Makse H A 2014 Avoiding catastrophic failure in correlated networks of networks *Nat. Phys.* **10** 762–7
[25] Leskovec J, Lang K J, Dasgupta A and Mahoney M W 2009 Community structure in large networks: natural cluster sizes and the absence of large well-defined clusters *Internet Math.* **6** 29–123
[26] Watts D J and Strogatz S H 1998 Collective dynamics of 'small-world' networks *Nature* **393** 440–2
[27] Newman M E J 2002 Assortative mixing in networks *Phys. Rev. Lett.* **89** 208701
[28] Catanzaro M, Boguñá M and Pastor-Satorras R 2005 Generation of uncorrelated random scale-free networks *Phys. Rev.* E **71** 027103
[29] Dorogovtsev S N, Mendes J F and Samukhin A N 2001 Size-dependent degree distribution of a scale-free growing network *Phys. Rev.* E **63** 062101
[30] Parshani R, Buldyrev S V and Havlin S 2011 Critical effect of dependency groups on the function of networks *Proc. Natl.Acad. Sci.* **108** 1007–10
[31] Guimerà R, Diaz-Guilera A, Vega-Redondo F, Cabrales A and Arenas A 2002 Optimal network topologies for local search with congestion *Phys. Rev. Lett.* **89** 248701
[32] Yan G, Zhou T, Hu B, Fu Z-Q and Wang B-H 2006 Efficient routing on complex networks *Phys. Rev.* E **73** 046108
[33] Freeman L C 1979 Centrality in social networks conceptual clarification *Social Networks* **1** 215–39
[34] Tarjan R 1972 Depth-first search and linear graph algorithms *SIAM J. Comput.* **1** 146–60