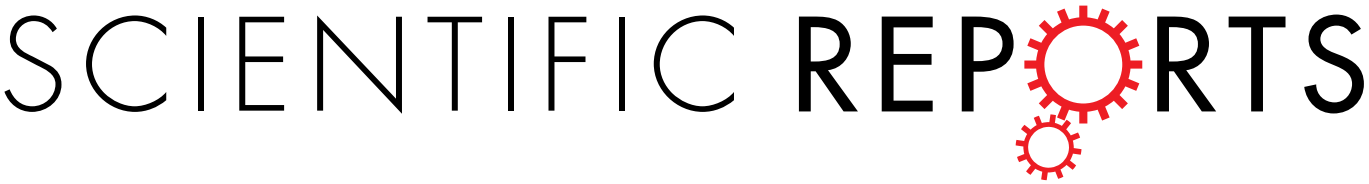

OPEN

Spreading to localized targets in complex networks

Ye Sun, Long Ma, An Zeng & Wen-Xu Wang

Received: 27 April 2016
Accepted: 14 November 2016
Published: 14 December 2016

As an important type of dynamics on complex networks, spreading is widely used to model many real processes such as the epidemic contagion and information propagation. One of the most significant research questions in spreading is to rank the spreading ability of nodes in the network. To this end, substantial effort has been made and a variety of effective methods have been proposed. These methods usually define the spreading ability of a node as the number of finally infected nodes given that the spreading is initialized from the node. However, in many real cases such as advertising and news propagation, the spreading only aims to cover a specific group of nodes. Therefore, it is necessary to study the spreading ability of nodes towards localized targets in complex networks. In this paper, we propose a reversed local path algorithm for this problem. Simulation results show that our method outperforms the existing methods in identifying the influential nodes with respect to these localized targets. Moreover, the influential spreaders identified by our method can effectively avoid infecting the non-target nodes in the spreading process.

Spreading is a fundamental dynamical process in real systems. It has been intensively studied in many different fields including physics, chemistry, social science, biology and computer science[1]. The reason behind this is that the emergence of many complex and heterogeneous connectivity patterns in a wide range of biological and social systems can be modeled and investigated by the spreading process in complex networks[2–4]. Examples include the epidemic contagion[5] and rumor/news propagation[6,7]. After more than a decade of study, our understanding on the properties of spreading processes in complex networks is now much deeper. Results are fruitful. For instance, the spreading on complex networks is found to undergo a second-order phase transition in most cases but could be explosive in synergistic epidemics[8,9], and the critical infection probability can be estimated by the mean-field theory[10]. The networks with heterogeneous degree distribution in general have a lower critical infection probability than those with homogeneous degree distribution[11]. The spreading records have also been applied to reconstruct the propagation networks[12]. In addition, some methods have been developed to predict the spreading coverage[13,14] and the predictability of the spreading has been discussed[15,16]. For a very recent comprehensive review, see ref. 10.

Recently, a large amount of attention has been paid to investigate the spreading ability of nodes in complex networks. Identification of the influential spreaders can, for example, help to design a better advertising strategy and a more efficient immunization strategy[17–21]. The traditional centrality measures can be naturally applied for this problem. In a pioneer paper[22], the authors pointed out that the k-shell methods can significantly outperform the traditional centralities such as degree[23] and betweenness[24]. After this work, a series of methods have been proposed[25,26]. For instance, the mixed degree decomposition method consider both the residual degree and the exhausted degree when decomposing the network and rank the nodes accordingly[27]; the iterative resource allocation method incorporates the centrality information of neighbors in ranking spreaders[28]; the path diversity has also been introduced to design the ranking method[29]. When spreading starts from multiple origins, the set of nodes with high spreading ability is not easy to find. So far, a number of papers have been devoted to solve this problem[30,31].

Despite the fact that the existing works on influential spreaders have greatly deepened our understanding of the spreading process in the microscopic level and led to many useful algorithms, one of the key problems is still overlooked, i.e. what would happen if the spreading process does not aim for all the nodes but only suppose to infect a small number of localized target nodes. This is an important research question from both theoretical and practical points of view. In recent literature, the problem of localized targets has been intensively studied by many researchers and was found to be very different from the global targets problem[10]. Examples include the target control of complex networks[32] and localized attack on networks[33]. The target spreading problem is actually inspired

School of Systems Science, Beijing Normal University, Beijing 100875, P. R. China. Correspondence and requests for materials should be addressed to A.Z. (email: anzeng@bnu.edu.cn)

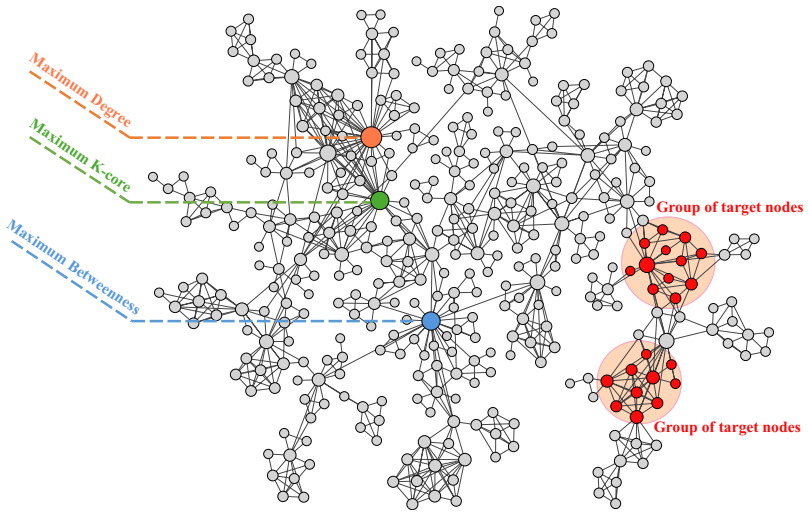

Figure 1. Illustration of the problem of spreading towards localized targets in complex networks. The network is the collaboration network of researchers working in network science (379 nodes and 914 links)[34,35]. The pink nodes are the targets that we want to infect. The high centrality nodes are respectively highlighted.

by many real cases. For instance, in online social networks (e.g. Facebook and Twitter), there is a great deal of information (e.g. advertisements, notifications and news) propagating between users. When the information aims to be sent to a specific group of users (namely the target users) instead of all users, this can be regarded as the target spreading problem. Specifically, in advertising based on online social networks, the beer advertisement should spread as much as possible to the potential adult customers but avoid propagating to teenagers. Another example is when there is a news about a job opening (post-doc or faculty position) in a university's physics department, the target nodes are the PhD students in physics field. However, if the hiring committee does not know the contact information about these PhD students but the email address of their collaborators from the published papers (i.e. the corresponding author of that paper), then it is essentially a target spreading problem in collaboration networks where one has to identify the best node that can propagate such information to most of these PhD students.

In this paper, we investigate the spreading ability of nodes towards localized targets in complex networks. we find that the existing methods for detecting influential spreaders all work poorly in this problem. We thus propose a reversed local path (RLP) algorithm which ranks the spreading ability of nodes by computing the local paths from the target nodes to other nodes. The method is validated with both artificial networks and real networks. The results show that our method can remarkably outperform the existing methods such as degree, k-shell and betweenness in identifying the nodes with high spreading ability towards the localized targets. Moreover, the influential spreaders identified by our method can effectively avoid infecting the non-target nodes in the spreading process. Besides the effectiveness, our method has advantage in the computational complexity compared to the existing methods. Though we consider the classic Susceptible-Infected-Recovered (SIR) model[1] in this paper, we believe that our method also works well in other spreading models and will have many practical applications in real systems.

Results

Spreading with localized targets. We first briefly describe the problem of spreading towards localized targets in complex networks. We consider a real network (e.g. the collaboration network of researchers working in network science) as shown in Fig. 1. Two groups of pink nodes are selected as the targeted nodes that we aim to infect. As the nodes in each group are well connected with each other, we call them localized targets. Besides these targets, the nodes with the highest degree, betweenness and k-shell values are also highlighted respectively. It is clear that these nodes are topologically far away from the target nodes, the virus or information starting from them has to pass through a lot of non-target nodes to reach the target nodes. If the infection probability is low, the spreading starting from these three nodes may even die out before reaching any of these target nodes. Therefore, the three nodes with highest centralities are no longer the best spreaders towards the localized targets.

We then quantitatively study the difference between the spreading with localized targets (i.e. a small group of nodes are targets) and globalized targets (i.e. all the nodes in the network are targets). To this end, we first define the spreading ability ρ_i of a node i as the fraction of infected target nodes given the spreading originated from node i. In this paper, we employ the SIR model to simulate the spreading process on networks. In the SIR model, an infected node makes contact and is able to transmit the disease with probability λ (called infection probability) to each of its neighbors. After infecting others, the infected node will become recovered and can never be infected again. Without loss of generality, we set the recovery probability $\mu = 1$. ρ_i can be obtained by simply computing the fraction of target nodes that are recovery nodes at the end given the spreading originated from node i. We first compute ρ_i of each node in Netsci network with 379 nodes and 914 links[34,35]. The dependence of ρ_i on the spreaders' degree in Netsci network with the globalized target case and the localized target case is shown in Fig. 2(a,b), respectively. In Fig. 2(a), i.e. the globalized target case, one can see that ρ_i strongly correlates with the spreaders' degree k_i. However, in the localized target case, the correlation between ρ and k is much weaker as shown in

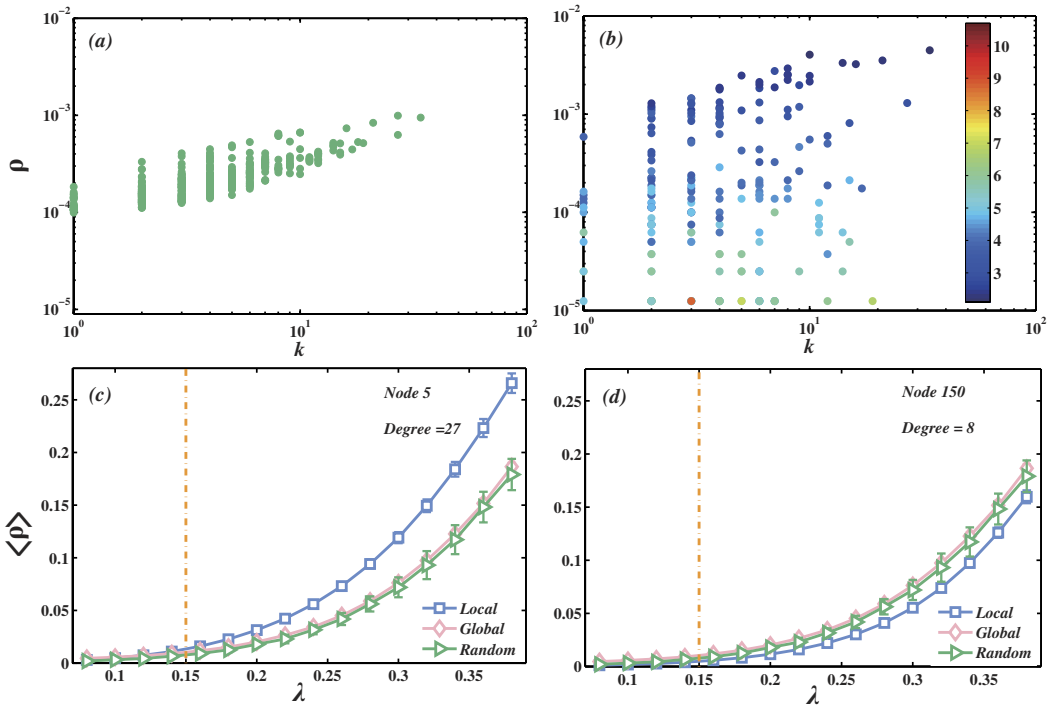

Figure 2. (**a**) The dependence of the fraction of infected target nodes ρ on the initial spreaders' degree k. In this sub-figure, all the nodes in the network are target nodes. (**b**) The dependence of the fraction of infected target nodes ρ on the initial spreaders' degree k and the mean shortest path length $\langle d\rangle$ from the spreader to the target nodes. The color of each point represents the $\langle d\rangle$ of the spreader. In this sub-figure, there are only 20 target nodes. A node is randomly selected as a center and the rest of the targets are placed in the nodes with the shortest path length no larger than 2 to the center. Center nodes are also target nodes. In both (**a**,**b**), the infection probability $\lambda = 0.12$, slightly smaller than the critical infection probability $\lambda_c = 0.15$. (**c**,**d**) The average fraction of infected target nodes $\langle\rho\rangle$ as a function of infection probability λ. In pink rhombus line, all the nodes in the network are target nodes. In green triangle line, we randomly select 20 nodes as the target nodes, while in blue square line, the method of choosing target nodes is the same as (**b**). The difference between (**c**) and (**d**) is that the center has $k = 27$ in (**c**) while $k = 8$ in (**d**). In all sub-figures, the networks are Netsci with $N = 379$ and $\langle k\rangle = 4.8$. The results are obtained by averaging 500 independent realizations.

Fig. 2(b). For a fixed degree, there is a wide spread of ρ values, which indicates that degree is no longer a good predictor of nodes' spreading ability. In Fig. 2(b), the color of each point represents the mean shortest path length $\langle d_i\rangle$ from the spreader i to the target nodes. One can see that the nodes with small $\langle d_i\rangle$ and large k_i tend to have high ρ_i.

To further understand above observations, we investigate the effect of different location of the targets in Fig. 2(c,d). We fix the number of target nodes as 20 and consider two scenarios, i.e. either the targets are randomly located in the network or they are located in a small area. To realize the second scenario, we first randomly pick up a node and set it as a center for this small area. This centre node is also considered to be one of the targets. The rest of the targets are placed in the nodes with the shortest path length not larger than 2 to the central node. We compare the average fraction of infected target nodes $\langle\rho\rangle$ as a function of the infection probability λ in these two scenarios. As a benchmark, we also plot $\langle\rho\rangle$ versus λ with the globalized targets in both Fig. 2(c) and (d). One can see that if the 20 targets are distributed randomly, the curve overlaps well with the curve of the globalized target case. However, when the targets are localized within two step distance, the $\langle\rho\rangle$ curve illustrates an apparent difference compared with the two cases above. These results also indicate that the localization of the targets makes the spreading properties significantly differs from the traditional case. The same conclusions can also be reached in Barabasi-Albert (BA) networks[36] with size $N = 500$ and mean degree $\langle k\rangle = 4$ (see Supplementary Information (SI)). In the following, we will mainly focus on how to accurately identify the node with high capability to spread the virus/information to the localized targets.

In this paper, we consider the cases where the target nodes cannot be chosen as seeds. This is a reasonable assumption supported by many real examples. For spreading a job news in the collaboration network, the contact information of a node is necessary if we want to select it as a seed. Unfortunately, the target nodes in this situation are young scholars (i.e. PhD students) whose contact information is usually unknown. In some other cases, the target nodes can still not be chosen as seeds even if their contact information is available. For instance, in the online social networks, the target nodes for a company are the potential buyers of its products. However, target users may refuse to send the advertisement of an unfamiliar product to their friends. The customers who have already used the products (i.e. non-target node) may forward the advertisement to their followers and help to promote the products.

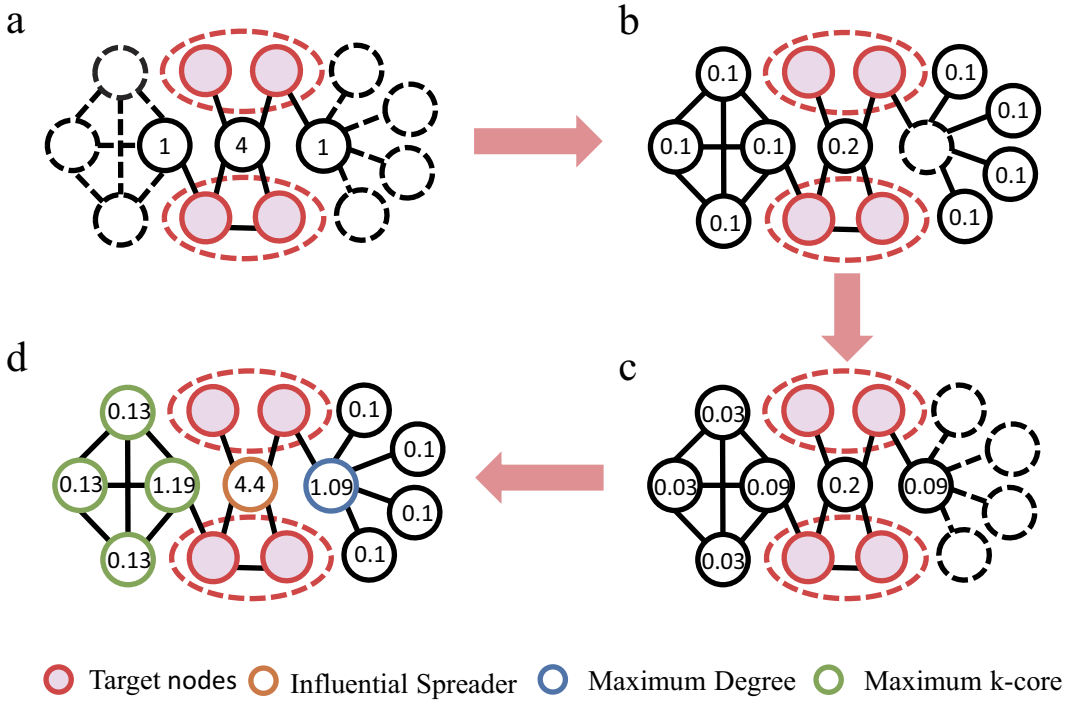

Figure 3. Illustrations of the reversed local path algorithm (RLP). The red nodes are target nodes and others are non-target nodes. (**a**) The nodes with numbers are the first-order neighbors of the targets. All irrelevant nodes and edges are marked in dashed lines. The numbers on the nodes are obtained by computing fA. (**b**) The nodes with numbers are the second-order neighbors of the targets. All possible paths with length 2 are considered and the numbers on the nodes are obtained by computing ϵfA^2. (**c**) The nodes with numbers are the third-order neighbors of the targets. All possible paths with length 3 are considered and the numbers on the nodes are obtained by computing $\epsilon^2 fA^3$. (**d**) The aggregated RLP score of non-target nodes are shown in this figure. The orange, blue and green nodes have maximum RLP, degree and k-core values, respectively.

The reversed local path method. In order to identify the spreaders that can easily infect the localized target nodes, we put forward a reversed local path (RLP) method. The basic idea for RLP is to compute the paths up to length 3 starting from the target nodes to other nodes. The paths with different lengths are aggregated to obtain the final score of a node. The nodes with large final score have high spreading ability towards the target nodes. The method is called reversed local path because only the relatively short paths are taken into account and the paths are counted in the opposite direction to the spreading process (i.e. calculation is from spreaders to target nodes in real spreading, but from target nodes to spreaders in RLP). Mathematically, the formula for RLP reads

$$S_{RLP} = \sum_{l=0}^{2} \epsilon^l f A^{l+1}, \tag{1}$$

where f is a $1 \times N$ vector in which the components corresponding to the target nodes are 1, and 0 otherwise. A is the $N \times N$ adjacency matrix of the network with $A_{ij} = 1$ indicating that node i connects to node j and $A_{ij} = 0$ otherwise. The product fA^{l+1} is an inner product. By definition, the score of nodes at a distance $l > 3$ from target nodes is zero. Here, ϵ is a tunable parameter controlling the weight of the paths with different lengths. In fact, the introduction of parameter ϵ is inspired by the well-known Katzs index[37]. Usually, ϵ is set to be a small value. We have tested different values of ϵ and find that there is an optimal ϵ for each network resulting in a maximum ranking accuracy (see SI). In this paper, we fix $\epsilon = 0.1$ which is near the optimal ϵ in many networks. We only take into account the paths with small length for the sake of efficiency[26]. We have checked that if we extend the path length to 10, the results will not be much better, sometimes even worse, depending on the setting of ϵ (see SI). In fact, the reversed computation (i.e. from target nodes to spreaders) can also significantly reduce the computational complexity, especially when the targets are few and the network is very large. The computational complexity to traverse the neighborhood of a node is simply the mean degree k of the network. If one estimates the spreading ability of each node by directly computing their local paths to target nodes, the computational complexity is $O(Nk^3)$ where N is the number of nodes in the network. However, with RLP the computational complexity can be reduced to $O(mk^3)$ where m is the number of the targets. As $m \ll N$ in the localized target problem, the RLP is much more efficient. The RLP process is illustrated with a toy network in Fig. 3. One can see that the most highly ranked node by RLP is different from the nodes with maximum degree and maximum k-shell. Besides, we also propose a simple linear model considering nodes' degree and average distance to targets (see SI). The results show that the

	Network properties				Random scheme				Local scheme	
Network	N	$\langle k\rangle$	D	λ_c	$\langle\tau\rangle_d$	$\langle\tau\rangle_b$	$\langle\tau\rangle_k$	$\langle\tau\rangle_{RLP}$	$\langle\tau\rangle_{LD}$	$\langle\tau\rangle_{RLP}$
Dolphins	62	5.13	8	0.172	0.776	0.531	0.775	**0.830**	0.642	**0.757**
Word	112	7.59	5	0.078	0.764	0.639	0.754	**0.815**	0.758	**0.821**
Jazz	198	27.70	6	0.027	0.665	0.519	0.671	**0.791**	0.633	**0.835**
E. coli	230	6.04	11	0.075	0.713	0.491	0.752	**0.840**	0.690	**0.833**
C. elegans	297	14.46	5	0.040	0.687	0.577	0.700	**0.780**	0.665	**0.780**
Netsci	379	4.82	17	0.142	0.443	0.305	0.415	**0.803**	0.629	**0.799**
Email	1133	9.62	8	0.057	0.759	0.637	0.775	**0.799**	0.718	**0.777**
Blog	1222	27.36	8	0.013	0.708	0.607	0.713	**0.724**	0.782	**0.792**
TAP	1373	9.95	12	0.065	0.675	0.352	0.669	**0.824**	0.526	**0.733**
Y2H	1458	2.67	19	0.163	0.301	0.289	0.346	**0.632**	0.586	**0.791**
HEP	5835	4.73	19	0.123	0.534	0.403	0.562	**0.641**	0.579	**0.708**
PGP	10680	4.55	24	0.056	0.494	0.357	0.509	**0.728**	0.517	**0.736**

Table 1. Structural properties and ranking results of different methods in real networks. Structural properties include network size (N), average degree ($\langle k\rangle$), network diameter (D), critical infection probability λ_c $\left(\text{which is computed by } \lambda_c = \frac{\langle k\rangle}{\langle k^2\rangle - \langle k\rangle}\right)$. The random scheme represents the case where 10% nodes are set as the target nodes and randomly distributed in the network. The local scheme stands for the case where 10% nodes are set as the target nodes and locate in the nodes with maximum distance L measured by the shortest path length to a randomly selected central node ($L=3$ in PGP, $L=4$ in HEP, $L=8$ in Y2H and $L=2$ in the rest of networks). According to Fig. 4, we compare τ of four methods including degree (τ_d), betweenness (τ_b), k-core (τ_k) and RLP (τ_{RLP}) in the random scheme. According to Fig. 5, we compare τ of two methods including Local degree (τ_{LD}) and RLP (τ_{RLP}). The infection probability for the SIR model in each network is set as λ_c. The results of the RLP method are highlighted.

linear combination of degree and average distance can indeed result in a higher accuracy. However, the results of RLP method are better than that of this linear model under different values of θ.

In spreading dynamics, several existing centrality indices are widely used to identify the influential spreaders in networks. The basic idea is that the spreading originated from the node with high centrality will finally reach more nodes. In this paper, we compare the RLP method with three existing representative centrality measures: degree[23], betweenness[24], k-shell[22] (See the Method section). Considering the findings in Fig. 2 that both degree and distance are essential factors affecting the spreading ability of nodes towards the localized targets, here we compare RLP with an additional index based on degree, called local degree (LD) method. In the LD method, the local degree of nodes with distance no larger than 3 to the target nodes is equal to their degree while the local degree of other nodes is zero (See the Method section).

Data and Metric. To validate the RLP method, we will apply it to both artificial and real networks. The artificial networks include the well-known Watts-strogatz (WS) model[38] and Barabasi-Albert (BA) model[36]. We also consider 10 real networks from both social and nonsocial systems. Social networks are: Dolphins (friendship)[39], Jazz (musical collaboration)[40], Netsci (collaboration network of network scientists)[34], Email (communication)[41], Blog (online blog network of politicians)[42]. Nonsocial networks are: Word (adjacency relation in English text)[34], *E. coli* (metabolic)[43,44], C. elegans (neural network)[45,46], TAP (yeast protein-protein binding network generated by tandem affinity purification experiments)[47,48], Y2H (yeast protein-protein binding network generated using yeast two hybridization)[49], HEP (collaboration network of high-energy physicists)[50], PGP (an encrypted communication network)[51]. Throughout this paper, we present the results of the two artificial networks and two selected real networks (i.e. Netsci and Y2H). The results of the other real networks are reported in Table 1.

For all the methods mentioned above, we generate the final ranking of nodes. In principle, a well-performing ranking algorithm should obtain a ranking as consistent as possible with the ranking based on nodes' spreading ability ρ. We then use the Kendall's tau rank correlation coefficient (τ)[52] to estimate how a certain obtained ranking is correlated to the ranking by the true spreading ability ρ of nodes (See the Method section). According to the definition of Kendall's tau coefficient, $-1 \le \tau \le 1$. In the most ideal case where $\tau = 1$, for each pair of two nodes i and j, if i is ranked higher than j by the method, the spreading originated from i will cover more targets than the spreading starting from j.

Simulation results. To begin our analysis, we first compare the accuracy τ of the above-mentioned ranking methods under different infection probability λ in Fig. 4. We consider the case where there are 30 randomly distributed targets in the network. Four networks are considered. In WS and BA networks (Fig. 4a,b), we do not show the results of the k-core method as the k-shell values of all the nodes in these two networks are almost the same. The results in each figure are obtained by averaging over 5000 independent realizations. The procedure is that we first take a realization of a network, investigate lots of target node sets in order to compute τ, and then average τ over many network realizations. However, for each of the real network cases (Fig. 4c,d), there is only one network and we just average the results over different target node sets. One immediate observation in Fig. 4 is that the RLP method has much higher accuracy τ than the other methods, especially when λ is small. However,

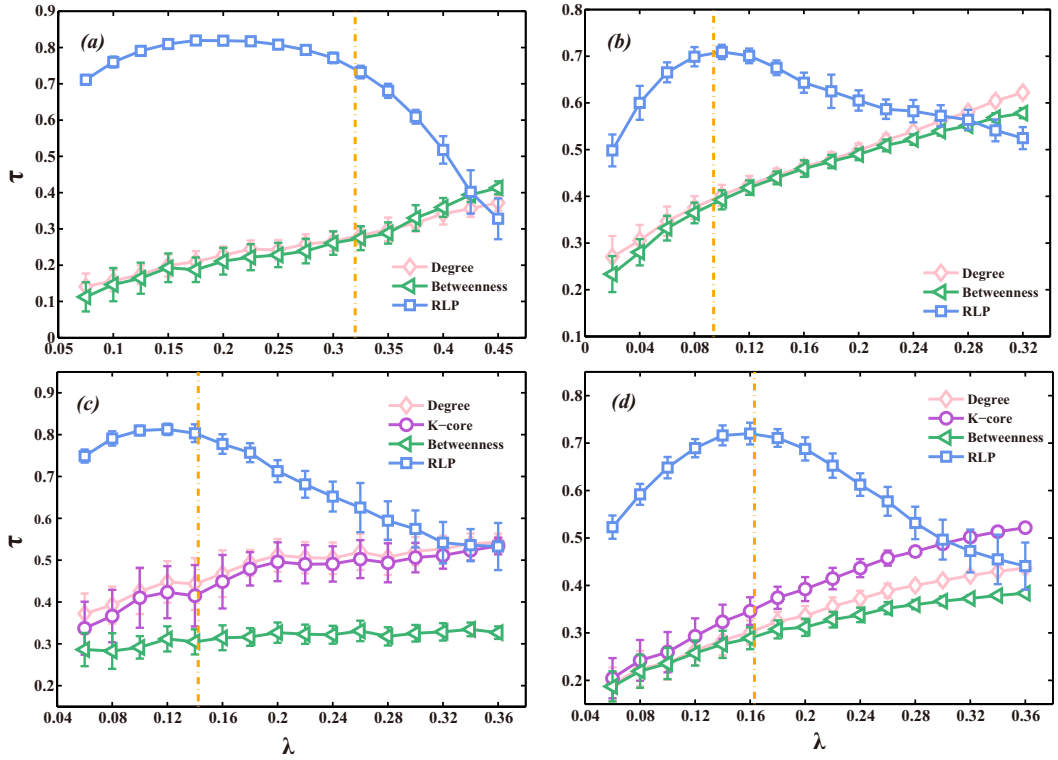

Figure 4. Kendall's tau rank correlation coefficient τ between the rankings obtained from different methods and the true spreading ability ρ under different infection probabilities λ. Four networks are considered, i.e. (**a**) WS, (**b**) BA, (**c**) Netsci and (**d**) Y2H networks. In each network, 30 target nodes randomly locate in the network. Ranking methods include degree (red diamonds), betweenness (green triangles), k-core (purple circles) or RLP (blue squares) methods. The orange dashed line corresponds to the critical infection probability. The results in each figure are obtained by averaging over 5000 independent realizations. In this figure, both WS and BA networks are with size $N = 500$ and mean degree $\langle k \rangle = 4$. The results of the artificial networks with bigger size can be found in SI.

when λ is too large and far exceeding the critical infection probability λ_c (marked by the orange vertical dashed lines in the figure), the spreading originated from each node may cover nearly the whole network including the target nodes. In this case, the final spreading coverage can no longer reflect the true spreading ability of nodes. Therefore, the τ value of RLP is similar to that of the other three methods when λ is large. Besides, we also compare the accuracy τ of RLP method and that of other centrality methods when all the nodes are target nodes in each network (see SI). The results show that the RLP method outperforms other centrality methods, especially when the infection probability is near the critical infection probability λ_c.

By computing the weighted paths up to distance 3 from the target nodes to other nodes, we are actually estimating the spreading influence from the target nodes to other nodes[26]. This is the inverse of the SIR dynamics, which is from nodes to targets. The parameter ϵ plays similar role of the infection probability λ. As the networks are undirected, the spreading influence from the nodes to targets can be approximated by that from targets to nodes. We can estimate the spreading ability of a node j by adding the estimated spreading influence from each target node to node j. Therefore, the RLP method works well, and also better than topological methods like centrality.

We then compare the performance of RLP and the local degree (LD) method in Fig. 5. The way we place the target nodes is the same as Fig. 2(b). We first select a node in the network as the so-called central node. There are m targets in the network and the $m-1$ targets randomly locate in the nodes with maximum distance L (measured by the shortest path length) to the central node. Apparently, when L is infinitely large, these m nodes distribute randomly in the network. The smaller L is, the more localized the targets are. Here, we set the value of the infection probability near the critical infection probability λ_c in each network. One can see that the RLP method constantly outperforms the LD method.

Figures 4 and 5 only show the results of τ in four networks. We further examine the performance of RLP and LD in the modeled networks with different sizes, the results show that that RLP can still significantly outperform LD when the network size is very large (as shown in SI). In addition, we applied our method to 12 real networks in this paper. The results of all these real networks are summarized in Table 1. One can see that in all the considered real networks, the RLP method outperforms the rest of other methods. In general, the advantage of RLP over other methods are larger in the real networks with high diameter D such as Netsci, Y2H, HEP and PGP. In these networks, the localized effect of the target nodes is stronger. To further verify this point, we study the effect of the community structure on our results. We consider the well-known GN-benchmark network model[53] and find that

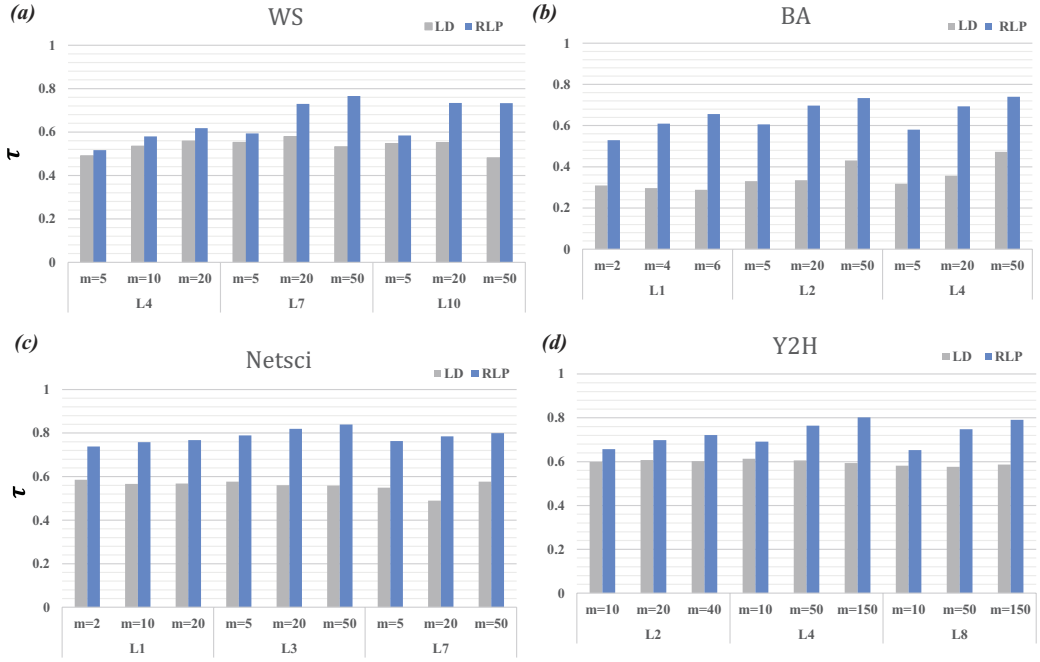

Figure 5. The spreading ability ranking accuracy τ under different m and L in four networks. The parameters for WS and BA networks are $N = 500$ and $k = 4$. The results in this figure are obtained by averaging over 5000 independent realizations.

as the community structure becomes more obvious, the accuracy of traditional centrality index tends to decrease while the accuracy of RLP tends to increases (see the results in SI). These results confirm that the target spreading problem in general becomes more challenging when the network diameter is large.

In fact, when we try to infect target nodes, some non-target nodes are infected as well. However, in many real systems the propagation towards the non-target nodes should be avoided. For instance, in advertising the beer company should avoid showing their advertisement to the kids when they try to promote their beer sale by posting advertisements in the online social networks. Accordingly, we investigate the ability of the RLP method in avoiding infecting non-target nodes and compare the results with the degree and LD methods. For each method, we pick up the most highly ranked node i. Given the spreading initialized from i, the fraction of finally infected target nodes and non-target nodes are respectively denoted as ρ_i and ν_i. In Fig. 6, we show the relation between ρ and ν under different infection probabilities when the three ranking methods are applied to four networks. The faster ρ increases with ν, the better the method is in avoiding infecting the non-target nodes. Clearly, the RLP method outperforms the degree and LD methods as it can achieve a high ρ with a very small ν. The advantage of RLP is smaller in BA network. This is because the network has one or several hub nodes (nodes with very large degree) and they are very easy to be infected. Once a hub node is infected, many neighboring non-target nodes will be easily infected. Though some other real networks have hub nodes too, these real networks have some level of community structure (like Fig. 1) such that the network diameter is large and the target nodes can form a local structure that is far away from the hub nodes.

Discussion

Identification of the influential spreaders is a very important problem from both theoretical and practical point of view. Though a number of methods have been proposed in the literature, the basic assumption for these works is that the spreading aims to infect all the nodes. Inspired by the fact that in many real systems only a small number of nodes in the network are intended to be infected, we put forward a target-oriented spreading problem in this paper. We find that this problem is significantly different from the traditional spreading problem in terms of the influential spreader identification. Specifically, the traditional centrality methods such as degree, betweenness and k-shell are found to be inefficient in finding the spreader that can effectively infect the target nodes. We thus propose a reversed local path method to rank the spreading ability of the nodes towards the target nodes. The simulation results indicate that our method can remarkably outperform the traditional methods, especially when the target nodes are relatively few and strongly localized. The methods are validated in both artificial and real networks. Finally, our method is found to be able to effectively suppress the infection to the non-target nodes.

In fact, the target spreading problem is closely related to the research topic on controlling complex networks which has been intensively investigated in recent years[32,54–57]. Both problems aim to affect a specific group of nodes in a network (either to propagate some information to them or to drive them to a desirable state). However, there are some key differences between these two problems that hinder the direct application of the approaches on network controllability to target spreading. The network controllability problem is formalized by a differential equation which usually relies on maximum matching algorithms to identify the driver nodes[54]. The spreading

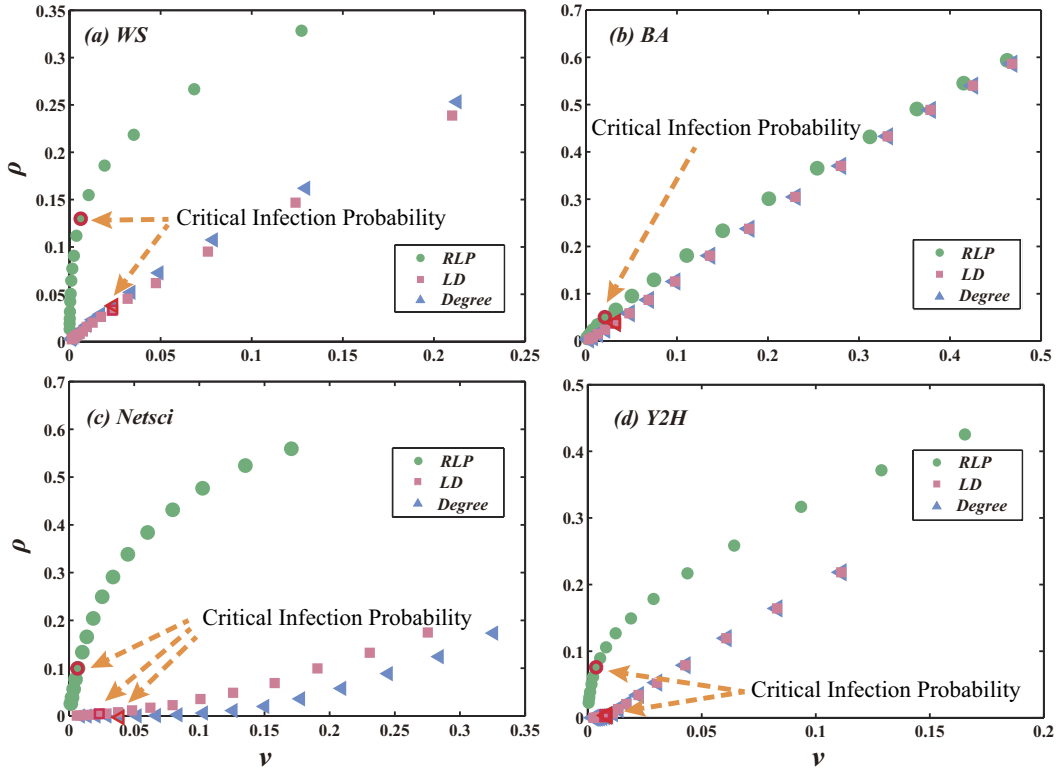

Figure 6. The relation between the fraction of infected target nodes ρ and the fraction of infected non-target nodes v under different infection probabilities when the RLP, LD and degree methods are applied to four networks. Each point in this figure represents the result obtained with a certain infection probability. The point corresponding to the critical infection probability is marked in the figure. In each network, 30 target nodes randomly locate within distance L to a center node. In WS network, $L=4$ and the center node has degree 4. In BA network, $L=2$ and the central node has degree 7. In Netsci network, $L=2$ and the central node has degree 19. In Y2H network, $L=2$ and the central node has degree 5. The network parameters for BA and WS are $N=500$ and $\langle k\rangle=4$. The results are obtained by averaging over 100 independent realizations.

problem, however, is described by some stochastic models such as SIR and SIS, and the influential seeds are identified usually by centrality metrics[24,58]. There are actually already many existing papers showing the difference between these two research problems[59,60]. For instance, ref. 59 reveals the low overlap between the high centrality nodes and the driver nodes for controlling networks.

We believe that this paper proposes a very general research problem and many related issues could be studied in the near future. For instance, to better understand the statistical properties of the target-oriented spreading process, one can systematically investigate the effect of target number and the topological distribution of the targets on the epidemic phase transition and the critical infection probability. Moreover, the method in this paper aims to maximize the coverage of the target nodes, a better method could try to maximize this objective and minimize the coverage of the non-target nodes simultaneously. Our method is based on the local paths, a better method might be designed based on the likelihood maximization approach[61]. In this way, not only a more accurate method could be developed, some theoretical estimation of the final infected nodes given the spreading originated from different nodes could be obtained. Finally, how to control the spreading process towards the target nodes while the virus or information is already propagation in the networks is also a meaningful research issue. We believe that our work serves as a very good starting point for these problems.

Methods

Existing Centrality indices. There are many existing centrality indices that can be used to identify the influential spreaders in networks. In this paper, we compare our method with three existing representative centrality measures.

(i) Degree centrality. The degree[23] of node i can be defined as $k(i)=\sum_{j\in G}a_{ij}$ where a_{ij} is a component of the network's adjacency matrix. Degree represents the number of neighbors this node has, which reflects the direct influence of this node to others.

(ii) Betweenness centrality. The betweenness centrality of node i, b_i, is defined as follows[24]. Between every combination of nodes a and b excluding i, we can obtain at least one shortest path. After respectively defining the number of all these paths and the paths though node i to be $n_{a,b}$ and $n_{a,b}(i)$, b_i is then given by:

$$b_i = \sum_{(a,b)} \frac{n_{a,b}(i)}{n_{a,b}} \tag{2}$$

(iii) k-shell decomposition. By removing nodes with degree less than or equal to k iteratively, the k-shell (also called k-core) method tends to have lower implementation complexity than betweenness and higher accuracy than both degree and betweenness[22]. The definite operations are as follows: We start by removing nodes with degree $k=1$ until there is no node left with $k=1$ in the network. Then the k-shell value of these removed nodes is set as $ks=1$. In step n, one should continually remove nodes with residual degree no more than n. According to the above operation, the nodes removed in step n have a k-shell value $ks=n$.

Local degree. Considering the findings in Fig. 2 that both degree and distance are essential factors affecting the spreading ability of nodes towards the localized targets, here we consider an additional index based on degree, called local degree (LD). Mathematically, the local degree kl of node i is given by:

$$kl_i = \begin{cases} \sum_j A_{ij}, & i \in \Omega \\ 0, & i \notin \Omega \end{cases} \tag{3}$$

where Ω is the node set including nodes within the distance $l=3$ from the target nodes.

Ranking accuracy. We use the Kendall's tau rank correlation coefficient (τ)[52] to estimate how a certain obtained ranking is correlated to the ranking by the true spreading ability ρ of nodes. The Kendall's tau coefficient considers a set of observations of the joint variables X and Y (in our case, X can be nodes' scores assigned by the ranking method and Y can be the spreading results ρ of all nodes). The tau value can be computed as

$$\tau = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} \mathrm{sgn}[(\mathrm{x_i} - \mathrm{x_j})(\mathrm{y_i} - \mathrm{y_j})]}{N(N-1)}, \tag{4}$$

where sgn(x) is the sign function, which returns 1 if $x>0$; -1 if $x<0$; and 0 for $x=0$. Here $(x_i-x_j)(y_i-y_j)>0$ means concordant, and negative means discordant.

References

1. Dorogovtsev, S. N., Goltsev, A. V. & Mendes, J. F. Critical phenomena in complex networks. *Rev. Mod. Phys.* **80,** 1275–1335 (2008).
2. Barrat, A., Barthelemy, M. & Vespignani, A. *Dynamical processes on complex networks* (Cambridge Univesity Press, 2008).
3. Kleineberg, K. K. & Boguñá, M. Evolution of the digital society reveals balance between viral and mass media influence. *Phys. Rev. X* **4,** 031046 (2014).
4. Del Vicario, M. *et al.* The spreading of misinformation online. *Proc. Natl. Acad. Sci. USA* **113,** 554–559 (2016).
5. Keeling, M. J. & Eames, K. T. Networks and epidemic models. *J. R. Soc. Interface* **2,** 295–307 (2005).
6. Pastor-Satorras, R. & Vespignani, A. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.* **86,** 3200–3203 (2001).
7. Moreno, Y., Nekovee, M. & Pacheco, A. F. Dynamics of rumor spreading in complex networks. *Phys. Rev. E* **69,** 066130 (2004).
8. Gomez-Gardenes, J., Lotero, L., Taraskin, S. N. & Prez-Reche, F. J. Explosive Contagion in Networks. *Sci. Rep.* **6,** 19767 (2016).
9. Cai, W., Chen, L., Ghanbarnejad, F. & Grassberger, P. Avalanche outbreaks emerging in cooperative contagions. *Nat. Phys.* **11,** 936–940 (2015).
10. Pastor-Satorras, R., Castellano, C., Van Mieghem, P. & Vespignani, A. Epidemic processes in complex networks. *Rev. Mod. Phys.* **87,** 925–979 (2015).
11. Moreno, Y., Pastor-Satorras, R. & Vespignani, A. Epidemic outbreaks in complex heterogeneous networks. *Europhys. Lett.* **26,** 521–529 (2002).
12. Shen, Z., Wang, W. X., Fan, Y., Di, Z. & Lai, Y. C. Reconstructing propagation networks with natural diversity and identifying hidden sources. *Nat. Commun.* **5,** 4323 (2014).
13. Colizza, V., Barrat, A., Barthélemy, M. & Vespignani, A. The modeling of global epidemics: Stochastic dynamics and predictability. *B. Math. Biol.* **68,** 1893–1921 (2006).
14. Wang, L. & Li, X. Spatial epidemiology of networked metapopulation: an overview. *Chin. Sci. Bull.* **59,** 3511–3522 (2014).
15. Holme, P. & Takaguchi, T. Time evolution of predictability of epidemics on networks. *Phys. Rev. E* **91,** 042811 (2015).
16. Pérez-Reche, F. J., Neri, F. M., Taraskin, S. N. & Gilligan, C. A. Prediction of invasion from the early stage of an epidemic. *J. R. Soc. Interface* **9,** 2085–2096 (2012).
17. Chen, Y., Paul, G., Havlin, S., Liljeros, F. & Stanley, H. E. Finding a better immunization strategy. *Phys. Rev. Lett.* **101,** 058701 (2008).
18. Schneider, C. M., Mihaljev, T., Havlin, S. & Herrmann, H. J. Suppressing epidemics with a limited amount of immunization units. *Phys. Rev. E* **84,** 061911 (2011).
19. Morone, F. & Makse, H. A. Influence maximization in complex networks through optimal percolation. *Nature* **524,** 65–68 (2015).
20. 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).
21. Pei, S. & Makse, H. A. Spreading dynamics in complex networks. *J. Stat. Mech.* **12,** P12002 (2013).
22. Kitsak, M. *et al.* Identification of influential spreaders in complex networks. *Nat. Phys.* **6,** 888–893 (2010).
23. Anderson, R. M., May, R. M. & Anderson, B. *Infectious diseases of humans: dynamics and control* (Oxford Univ. Press, Oxford, UK, 1991).
24. Freeman, L. C. A set of measures of centrality based on betweenness. *Sociometry* **40,** 35–41 (1977).
25. Hébert-Dufresne, L., Allard, A., Young, J. G. & Dubé, L. J. Global efficiency of local immunization on complex networks. *Sci. Rep.* **3,** 2171 (2013).
26. Bauer, F. & Lizier, J. T. Identifying influential spreaders and efficiently estimating infection numbers in epidemic models: A walk counting approach. *Europhys. Lett.* **99,** 68007 (2012).
27. Zeng, A. & Zhang, C. J. Ranking spreaders by decomposing complex networks. *Physi. Lett. A* **377,** 1031–1035 (2013).
28. Ren, Z. M., Zeng, A., Chen, D. B., Liao, H. & Liu, J. G. Iterative resource allocation for ranking spreaders in complex networks. *Europhys. Lett.* **106,** 48005 (2014).

29. Chen, D. B., Xiao, R., Zeng, A. & Zhang, Y. C. Path diversity improves the identification of influential spreaders. *Europhys. Lett.* **104,** 68006 (2013).
30. Zhao, X. Y., Huang, B., Tang, M., Zhang, H. F. & Chen, D. B. Identifying effective multiple spreaders by coloring complex networks. *Europhys. Lett.* **108,** 68005 (2014).
31. Shuai, X., Ding, Y. & Busemeyer, J. Multiple spreaders affect the indirect influence on Twitter. In: *Proceedings of the 21st international conference companion on World Wide Web. ACM* 597-598 (2012).
32. Gao, J., Liu, Y. Y., D'Souza, R. M. & Barabási, A. L. Target control of complex networks. *Nat. Commun.* **5,** 5415 (2014).
33. Holme, P., Kim, B. J., Yoon, C. N. & Han, S. K. Attack vulnerability of complex networks. *Phys. Rev. E* **65,** 056109 (2002).
34. Newman, M. E. J. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E* **80,** 036104 (2006).
35. Newman, M. E. J. Community centrality. http://www-personal.umich.edu/mejn/centrality/ Date of access: 11/07/2014 (2006).
36. Barabási, A. L. & Albert, R. Emergence of scaling in random networks. *Science* **80,** 509–512 (1999).
37. Katz, L. A new status index derived from sociometric analysis. *Psychometrika* **80,** 39–43 (1953).
38. Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature* **80,** 440–442 (1998).
39. Lusseau, D. *et al.* The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. *Behav. Ecol. Sociobiol.* **80,** 396–405 (2003).
40. Gleiser, P. M. & Danon, L. Community structure in jazz. *Advs. Complex Syst.* **80,** 565–573 (2003).
41. Guimera, R., Danon, L., Diaz-Guilera, A., Giralt, F. & Arenas, A. Self-similar community structure in a network of human interactions. *Phys. Rev. E* **80,** 065103 (2003).
42. Adamic, L. A. & Glance, N. The political blogosphere and the 2004 US election: divided they blog. In: *Proceedings of the 3rd international workshop on Link discovery. ACM* **80**, 651-654 (2005).
43. Overbeek, R. *et al.* WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids. Res.* **28,** 123–125 (2000).
44. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabási, A. L. The large-scale organization of metabolic networks. *Nature* **80,** 651–654 (2000).
45. Jeong, H. *et al.* The large-scale organization of metabolic networks. *Nature* **407,** 651–654 (2000).
46. Duch, J. & Arenas, A. Community detection in complex networks using extremal optimization. *Phys. Rev. E* **80,** 027104 (2005).
47. Rigaut, G. *et al.* A generic protein purification method for protein complex characterization and proteome exploration. *Nat. Biotechnol.* **17,** 1030–1032 (1999).
48. Gavin, A. C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **80,** 141–147 (2002).
49. Jeong, H., Mason, S. P., Barabási, A. L. & Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **80,** 41–42 (2001).
50. Newman, M. E. J. The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci.* **98,** 404–409 (2001).
51. Boguna, M., Pastor-Satorras, R., Diaz-Guilera, A. & Arenas, A. Models of social networks based on social distance attachment. *Phys. Rev. E* **70,** 056122 (2004).
52. Kendall, M. G. A new measure of rank correlation. *Biometrika* **80,** 81–93 (1938).
53. Girvan, M. & Newman, M. E. J. Community structure in social and biological networks. *Proc. Natl. Acad. Sci.* **99,** 7821–7826 (2002).
54. Liu, Y., Slotine, J. J. & Barabási, A. L. Controllability of complex networks. *Nature* **473,** 167–73 (2011).
55. Yuan, Z., Zhao, C., Di, Z. R., Wang, W. X. & Lai, Y. C. Exact controllability of complex networks. *Nat. Commun.* **4,** 2447 (2013).
56. Yan, G., Ren, J., Lai, Y. C., Lai, C. H. & Li, B. Controlling complex networks: how much energy is needed? *Phys. Rev. Lett.* **108,** 218703 (2012).
57. Jia, T. *et al.* Emergence of bimodality in controlling complex networks. *Nat. Commun.* **4,** 2002 (2013).
58. Sabidussi, G. The centrality index of a graph. *Psychometrika* **31,** 581–603 (1966).
59. Galbiati, M., Delpini, D. & Battiston, S. The power to control. *Nat. Phys.* **9,** 126–128 (2013).
60. Menichetti, G., DallAsta, L. & Bianconi, G. Network controllability is determined by the density of low in-degree and out-degree nodes. *Phys. Rev. Lett.* **113,** 078701 (2014).
61. Altarelli, F., Braunstein, A., Dall'Asta, L., Lage-Castellanos, A. & Zecchina, R. Bayesian inference of epidemics on networks via belief propagation. *Phys. Rev. Lett.* **112,** 118701 (2014).

Acknowledgements

We thank an anonymous reviewer for helpful suggestions which improve this paper. This work is supported by the National Natural Science Foundation of China (Nos 61603046 and 11547188), Natural Science Foundation of Beijing (No. 16L00077) and the Young Scholar Program of Beijing Normal University (No. 2014NT38).

Author Contributions

A.Z. designed the research, Y.S. and L.M. performed the numerical simulations, all the authors analyzed the results and wrote the paper.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Sun, Y. *et al.* Spreading to localized targets in complex networks. *Sci. Rep.* **6**, 38865; doi: 10.1038/srep38865 (2016).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.