Physics Letters A 380 (2016) 837–842

Contents lists available at ScienceDirect

Physics Letters A

www.elsevier.com/locate/pla

Identifying multiple influential spreaders in term of the distance-based coloring

Lei Guo [a], Jian-Hong Lin [a], Qiang Guo [a], Jian-Guo Liu [a,b,*]

[a] Research Center of Complex Systems Science, University of Shanghai for Science and Technology, Shanghai 200093, People's Republic of China
[b] Data Science and Cloud Service Research Centre, Shanghai University of Finance and Economics, Shanghai 200433, People's Republic of China

ARTICLE INFO

Article history:
Received 12 October 2015
Received in revised form 24 December 2015
Accepted 28 December 2015
Available online 31 December 2015
Communicated by C.R. Doering

Keywords:
Node spreading influence
Multiple spreaders
Distance-based coloring

ABSTRACT

Identifying influential nodes is of significance for understanding the dynamics of information diffusion process in complex networks. In this paper, we present an improved distance-based coloring method to identify the multiple influential spreaders. In our method, each node is colored by a kind of color with the rule that the distance between initial nodes is close to the average distance of a network. When all nodes are colored, nodes with the same color are sorted into an independent set. Then we choose the nodes at the top positions of the ranking list according to their centralities. The experimental results for an artificial network and three empirical networks show that, comparing with the performance of traditional coloring method, the improvement ratio of our distance-based coloring method could reach 12.82%, 8.16%, 4.45%, 2.93% for the ER, Erdős, Polblogs and Routers networks respectively.

1. Introduction

Identifying multiple influential spreaders is of significance for understanding the physics of the diffusion process on complex networks [1–3]. For example, it is helpful for developing efficient strategy to hinder the diseases spreading [4]. And it is of importance for identifying the influential users to release the advertisement information [5,7].

So far, a lot of methods have been proposed to identify the influential spreaders in networks such as the degree centrality, betweenness centrality, closeness centrality and coreness centrality [8–15]. When multiple spreaders are considered simultaneously, the distance between multiple spreaders becomes a crucial parameter that determines the extent of the spreading [5]. Recently, Kitsak et al. [6] argued that the node spreading influence is determined by its location in a network. Hu et al. [16] explored the spreading influence of multiple spreaders in community networks and found that the community hubs can infect more nodes. Rodriguez and Laio [17] proposed an approach to identify multiple spreaders based on the idea that influential nodes can be characterized by a higher density and a relatively large distance from nodes with higher densities. This method can identify influential nodes regardless of network structures. However, for different threshold parameters, the multiple spreader node set always be different. Recently, when choosing multiple spreaders, Morone and Makse [18] proposed a method which takes the optimal percolation into consideration. The percolation-based method provides exact solutions of the maximization problem and presents a theoretical framework to identify the multiple spreaders [18]. Zhao et al. [19] introduced the graph coloring method, namely IS method, into the influential spreader identification and found that the performance could be enhanced greatly. Compared with the density-based method [17], the graph coloring method can be expressed accurately in a mathematical way and the result of the coloring method in a network is identical. For the traditional IS method, the distance between the influential spreaders is around two. However, when the number of multiple spreaders is small and the average distance in a network is large, the influential spreaders obtained by IS method stay together in a relatively small range, which can result in the bad performance of influential spreaders. As shown in subplot (b) of Fig. 1, the simulation result shows that the effect of the multiple spreaders can be improved when the distance between influential spreaders is larger. Therefore, it is a key question to analyze the effect of the distance among multiple spreaders for the graph coloring method.

In this paper, we present an improved coloring method, namely IIS method, to analyze the relationship between the distance of multiple spreaders and the spreading influence by increasing the distance between multiple spreaders. In our method, we color each node with the rule that nodes are colored whose distance between multiple nodes is at least r which is close to the average distance

* Corresponding author at: Research Center of Complex Systems Science, University of Shanghai for Science and Technology, Shanghai 200093, People's Republic of China.
E-mail address: liujg004@ustc.edu.cn (J.-G. Liu).

http://dx.doi.org/10.1016/j.physleta.2015.12.031
0375-9601/

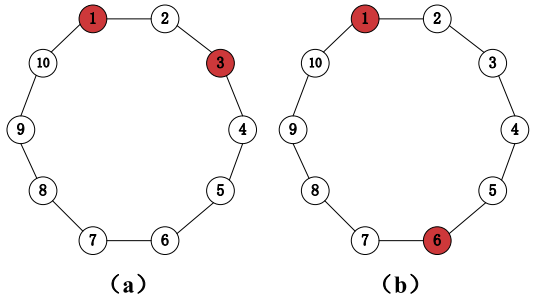

Fig. 1. (Color online.) An example network consists of 10 nodes and 10 links. As shown in the subplot (a), the initial two spreaders are 1 and 3 when the distance between initial spreaders is at least two. And as shown in the subplot (b), when the distance between initial spreaders is at least five, the initial two spreaders are 1 and 6. One can find that the effect of multiple spreaders can be improved when the distance between initial spreaders is taken into account.

of a network. When all nodes are colored, nodes with the same color are ranked according to their centralities and we choose the nodes at the top positions of the ranking list as multiple spreaders in the color set where the node with maximum degree is located. By implementing the method for an artificial network and three empirical networks, we find out that the performance of the IIS method is better than the ones of traditional IS method when the number of multiple spreaders is small and the value of r is close to the average distance of a network.

2. Methods

2.1. The traditional coloring method

Normally, a network $G = (N, E)$ with N nodes and E links could be described by an adjacent matrix $\mathbf{A} = \{a_{ij}\}$ where $a_{ij} = 1$ if node i is connected by node j, and $a_{ij} = 0$ otherwise. As one of graph coloring problem theorems [20], the four-color theorem states that, given any plane graphs, no more than four colors are required to color the regions of the plane graph so that no two adjacent regions have the same color [21,22]. Combining the graph coloring problem with nodes' centralities, Zhao et al. [19] proposed the IS method to identify the influential spreaders. The main steps of the traditional IS method are as follows. Firstly, a network $G = (N, E)$ is colored with the rule that no two adjacent nodes have the same color [23]. When all the nodes are colored, each node in node set corresponds to a kind of color. Secondly, the nodes with the same color are classified into an independent set. Finally, the nodes at the top positions of the ranking list in a same set are chosen as multiple spreaders in the color set where the node with maximum degree is located.

2.2. The improved coloring method

In this paper, we present an improved distance-based coloring method, namely IIS method, to identify multiple influential spreaders. Firstly, we color a given network with the rule that the nodes are colored whose distance between nodes is at least r. Secondly, when all the nodes are colored, the nodes with same color are classified into a same set. When the sets are settled, the distance between nodes in a same set is at least r. Thirdly, we choose the nodes at the top positions of the ranking list in a same set as multiple spreaders according to their centralities in the color set where the node with maximum degree is located. In this paper, we take the statistical properties of networks into consideration and consider the condition that the value of r is close to the average distance d of a network, i.e. $r = 3$ and $r = 4$ for the ER, Erdős and Polblogs networks and $r = 4$ and $r = 5$ for the Routers network in the IIS method respectively.

In the IIS methods, we rank the node according to the degree centrality, betweenness centrality, closeness centrality and coreness centrality respectively. And then we color the network and choose the multiple spreaders. Taking the degree centrality as an example, the details of the IIS method are:

Step 1: According to the degree centrality, rank the node in descending order, such that $k(1) \geq k(2) \geq \cdots \geq k(n)$, where $k(i)$ denotes the degree of node i;

Step 2: Let $\pi(i) = m$, where $\pi(i)$ is a color function which denotes the color label of node i, m is a color label. And in the first iteration, $i = 1, m = 1$;

Step 3: Let $C(m) = \{i \mid \pi(i) = m\}$, where $C(m)$ is a set containing nodes with the same color label m. For an uncolored node j, if the distance between node j and nodes in $C(m)$ is at least r, then $\pi(j) = m$;

Step 4: Let $m := m + 1$, then choose a node at the top positions of the ranking list from the uncolored node set and back to step 3. The process will stop until all the nodes are colored.

The degree of a node i is defined as the number of its neighbors, namely

$$k_i = \sum_{j=1}^{N} a_{ij}, \tag{1}$$

where a_{ij} is the element of matrix $\mathbf{A}$. Degree centrality performs well in evaluating nodes' spreading influences. It is widely applied for its simplicity and low computational cost [12].

The betweenness centrality [10] of node i is defined as the fraction of shortest paths between node pairs that pass through the node i. The betweenness centrality of node i can be denoted by

$$BC_i = \sum_{s \neq i \neq t} \frac{n_{st}^i}{n_{st}}, \tag{2}$$

where n_{st} is the number of shortest paths between nodes s and t, and n_{st}^i denotes the number of shortest paths between s and t which pass through node i.

The closeness centrality [11] of node i is the reciprocal of the sum of distances to all other nodes of i, which can be defined as

$$CC_i = \sum_{j=1}^{N} \frac{N}{d_{ij}}, \tag{3}$$

where N is the number of nodes and d_{ij} is the distance between node i and node j.

The k-shell method decomposes a network into different k-core values and could be implemented to identify the network core [13]. The decomposition process can be stated as follows: First, all the nodes with degree $k = 1$ are removed until all nodes' degrees are larger than one. Then, the removed nodes form a node set whose k-core value equals to one. Thirdly, we repeat the removing process by the similar manner for the nodes with degree $k = 2$ to get the node set whose k-core value equals to two. This removing procedure is repeated until all nodes of the networks are removed and assigned a k-core value.

3. Experimental results

3.1. Data description

To evaluate the performance of the IIS method, we implement the IIS method for an artificial network and the three empirical

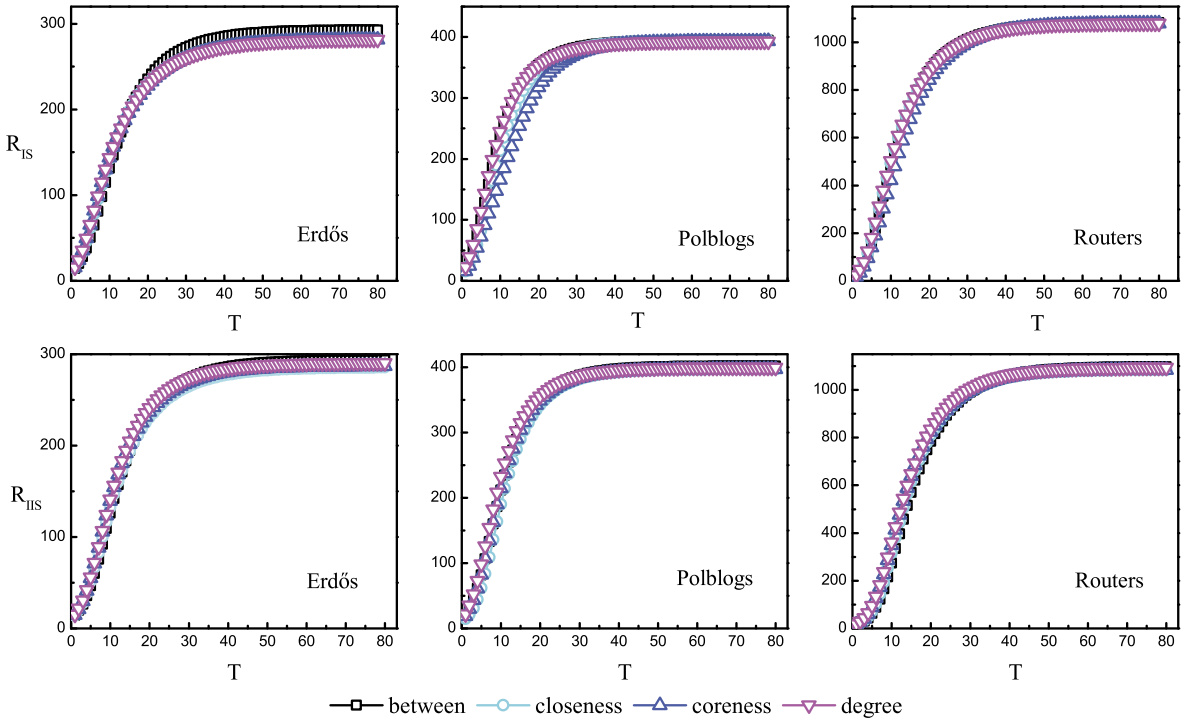

Fig. 2. (Color online.) The spreading influence is obtained by the SIR spreading process. The iteration time T varies from 1 to 80 when the number of spreaders is $n = 10$ and the transmission rate $\lambda = 0.4$. The results are averaged over 10,000 independent runs.

Table 1
Basic statistical features of the ER, Erdős, Polblogs and Routers networks, including the number of nodes N, the number of links E, the average degree $\langle k \rangle$ and the average distance d.

Network	N	E	$\langle k \rangle$	d
ER	10,000	64,963	12.91	3.79
Erdős	474	1639	6.92	3.83
Polblogs	643	2280	7.09	3.83
Routers	2113	6632	6.28	4.61

networks including the ER, Erdős, Polblogs and Routers networks. The ER network is consisted of 10,000 nodes and 64,963 links with edge probability $p = 0.0013$. The Erdős network is a scientific collaboration network. Each node represents the scientist whose Erdős number is 1 and the link represents the cooperative connection between each pair of scientists [13]. The Polblogs network is a communication relationship network. The node represents the owner of blogs and the link represents the communication. The Routers network is a technological network. The links represent the communication between different routers. The data sets can be available from the web site http://networkrepository.com/. The statistical properties of the artificial network and three empirical networks are shown in Table 1.

3.2. Measurement

In this paper, the nodes are ranked firstly according to the degree centrality, betweenness centrality, closeness centrality and coreness centrality respectively. And then we choose the multiple spreaders with the traditional IS method and the IIS method. Finally, we choose the multiple spreaders. When the multiple spreaders are settled, we use the susceptible-infected-recovered (SIR) [24] epidemic model to simulate the spreading process of multiple spreaders.

The SIR model is widely used to simulate the spreading process in networks. In SIR model, the nodes can be in one of the three states [25]: (i) Susceptible individuals represent the individuals who are easy to be infected; (ii) Infected individuals represent individuals who have been infected and are able to spread the disease to susceptible individuals; (iii) Recovered individuals represent individuals who have been recovered and will never be infected again. In each time step, the infected nodes start to infect their susceptible neighbors with the spreading rate β, and the infected node can become susceptible in one time step with the recovery rate μ. Finally, the number of nodes generated by the multiplely-infected node is denoted as its spreading influence. In this paper, we define the effective transmission rate $\lambda = \beta/\mu$ by fixing the recovery rate $\mu = 0.1$ [19] and all the results are averaged over 10,000 realizations and 50 time steps [26].

To compare the results of the IIS method and the traditional IS method, an improvement ratio Δr_R is utilized, which can be denoted by

$$\Delta r_R = \frac{R_{IIS} - R_{IS}}{R_{IS}} \times 100\%, \tag{4}$$

where R_{IIS} and R_{IS} denote the spreading influence of multiple spreaders obtained by the IIS method and IS method respectively when we use the SIR model to simulate the spreading process. Clearly, $\Delta r_R > 0$ indicates the advantage of the IIS method. And the larger the Δr_R is, the better the performance of the IIS method would be.

3.3. Numerical results

In this paper, we compare the performance of IIS method with the performance of IS method based on four centralities including the degree, closeness, betweenness and coreness.

Fig. 2 shows the spreading influence of the IIS method and IS method when the time step T in SIR model varies from 1 to 80. The results show that in the Erdős, Polblogs and Routers networks, each time step T is different when the spreading influence approaches a steady value. But one can find that the spreading influence approaches a steady value when the time step T is no

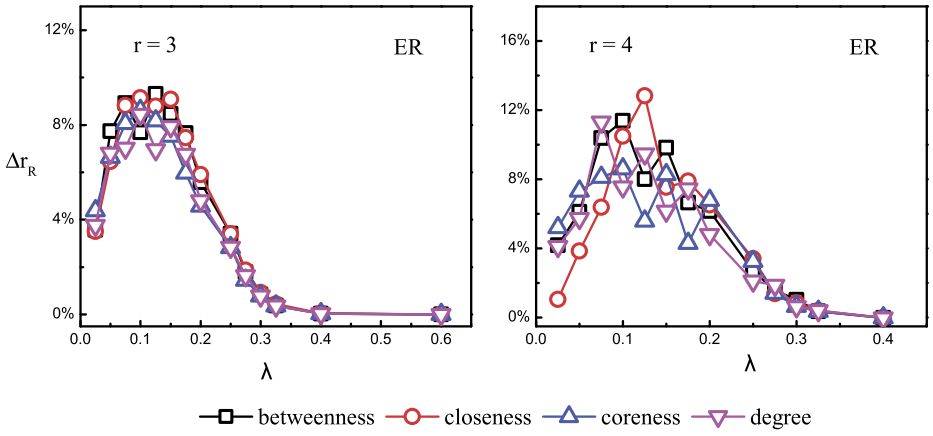

Fig. 3. (Color online.) The improvement ratio Δr_R to different parameter λ for the ER network where the number of multiple spreaders $n = 10$. The results are averaged over 10,000 independent runs.

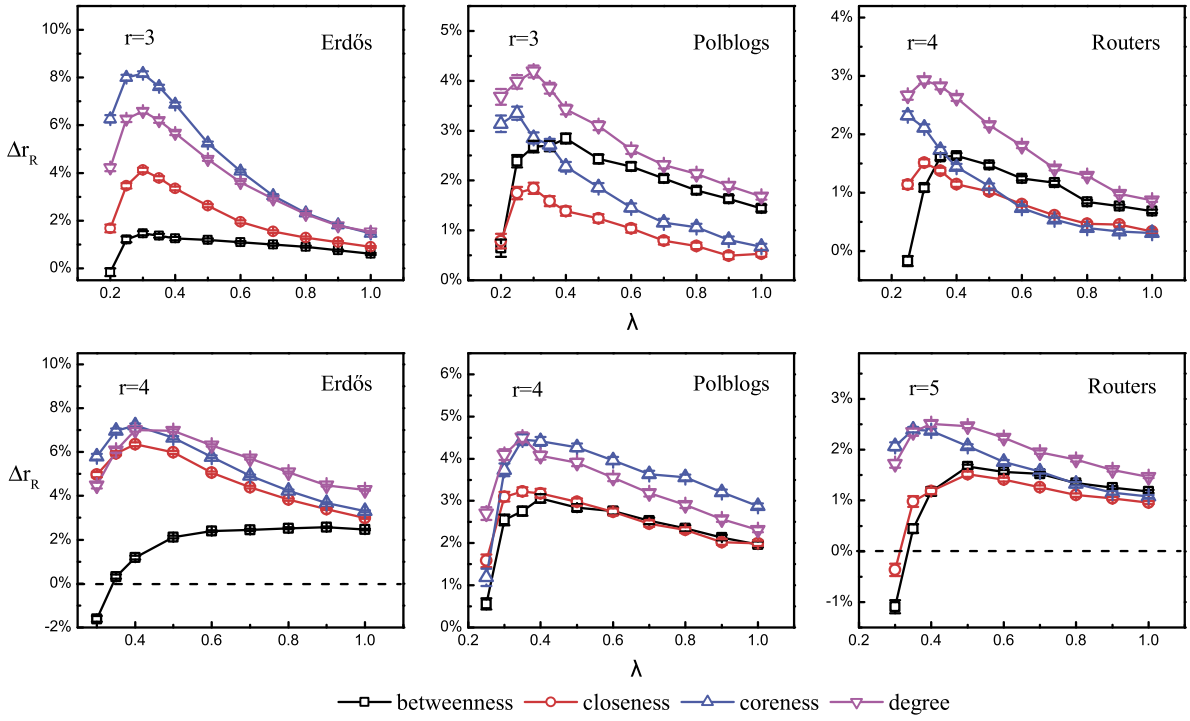

Fig. 4. (Color online.) The improvement ratio Δr_R to different parameter λ for the Erdős, Polblogs and Routers networks where the number of multiple spreaders $n = 20$. The results are averaged over 10,000 independent runs.

larger than 50. The IIS method and IS method can identify the influential nodes effectively when the time step $T = 50$. So we set the time step $T = 50$ in the SIR model.

As shown in Fig. 3, the improvement ratio Δr_R for the ER network is larger than 0, indicating that the IIS method performs better than the IS method. Specifically, the improvement ratio of the IIS method can reach 12.82% when $r = 4$, transmission rate is 0.125 and the centrality index is closeness. However, we can find that the improved ratio is close to 0 when the transmission rate becomes larger than 0.4. The main reason is that all the nodes in the ER network would be infected with large transmission rate. Figs. 4 and 5 report the improvement ratio of the IIS method based on the four centralities when the number of multiple spreaders is set as 20 and 10 for the Erdős, Polblogs and Routers networks respectively. As shown in Fig. 4, for most cases, the improvement ratio Δr_R is larger than 0, indicating that the IIS method performs better than the IS method when the number of multiple spreaders is 20. Specifically, in Erdős network, the improvement ratio could reach 8.16% when transmission rate is 0.3 and the centrality index is coreness. The similar results could be obtained in Polblogs and Routers networks. For example, when $r = 4$, transmission rate is 0.35 and the centrality index is degree, the improvement ratio Δr_R could reach 4.51%, 2.83% for Polblogs and Routers networks respectively. Furthermore, we investigate the performance of the IIS method when the number of the initial spreaders is 10 in Fig. 5. One can find that, the improvement ratio Δr_R could reach 5.19%, 2.27%, 1.22% for the Erdős, Polblogs and Routers networks when $r = 4$, transmission rate is 0.3 and the centrality index is degree. The improvement ratio with the standard deviation as error bar has been added in Fig. 4 and Fig. 5 respectively. The results show that the error bar becomes smaller with the transmission rate. However, there are some points in Fig. 3, Fig. 4 and Fig. 5 where the improvement ratio is below zero. The reason is that the node spreading influence is related to the spreading dynamics [27, 28], especially the transmission rate. In this paper, the distance between spreaders obtained by IIS method is larger than ones got-

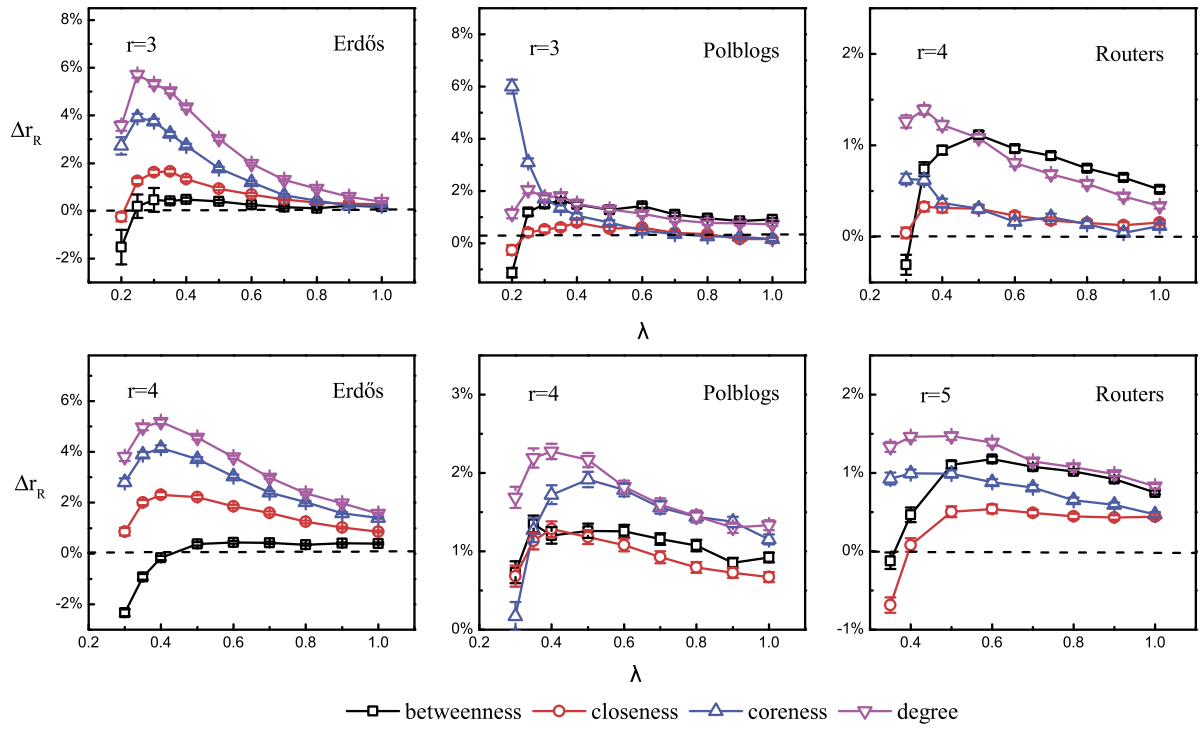

Fig. 5. (Color online.) The improvement ratio Δr_R to different parameter λ for the Erdős, Polblogs and Routers networks where the number of multiple spreaders $n = 10$. The results are averaged over 10,000 independent runs.

ten by the IS method. So when the transmission rate is small, the spreaders gotten by the IIS method can infect less nodes than the ones obtained by IS method, resulting in the condition that the improvement ratio is below zero.

The analysis for an artificial network and three empirical networks shows that the spreading influence of the multiple spreaders could be improved when the distance of multiple spreader is close to the average distance of the network. In addition, the results of the IIS method and the traditional IS method using four centrality indexes show that the IIS method can be further applied in other centrality indexes.

4. Conclusion and discussions

Identifying the influential spreaders is important for deeply understanding the diffusion process on networks. It is believed that within a certain range the more sufficiently dispersed the spreaders are, the better the spreading efficiency is [19]. In this paper, we propose an improved distance-based coloring method, namely IIS method, to enhance the performance of the multiple spreaders where the distance is taken into account. Firstly, the nodes are colored whose distance between nodes is r which is close to the distance of a network. And then nodes with the same color are classified into a set and finally we choose the nodes at the top positions of ranking list according to their centrality in the color set where the node with maximum degree is located. In this paper, the SIR model is introduced to simulate the spreading process, we choose the value of r that is close to the average distance d of a network. The results of an artificial network and three empirical networks show that the improvement ratio Δr_R is larger than 0. For example, the improvement ratio Δr_R of the Erdős network would reach the value 8.16% when the number of multiple spreaders $n = 20$ and the centrality index is coreness. The results of improvement ratio show that the performance of IIS method can perform better than the ones of traditional IS method for different transmission rate.

For the artificial network and three empirical networks, the results in Fig. 3, Fig. 4 and Fig. 5 show that there is always an optimal transmission rate λ under which the spreading influence can reach the maximum value and the optimal transmission rate λ becomes larger with the increase of distance between multiple spreaders. The spreading influence is related to the spreading dynamics [27,28], especially the transmission rate. For small transmission rate, the distance between the initial spreaders should be small. While for large transmission rate, the distance should be large for initial spreaders. So it would be a challenge work to analysis the relationship between the distance among spreaders and the transmission rate. In this paper, the value r is equal to the average distance of a network. But there are other properties of a network that may affect the performance of the method. For example, the sparsity of a network may affect the distance between spreaders and so the performance of IIS method may be affected as well. Furthermore, the assortativity coefficient affects the links between nodes. When the assortativity coefficient of a network is very high, which means there is high probability that important nodes stay together in a relative small range, the performance of the IIS method is affected. The future work would focus on how to analyze the relationship between the network structure and the optimal distance between multiple spreaders.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Nos. 71371125, 61374177, 71271126 and 71171136), the Doctoral Program of Higher Education (Grant No. 20120078110002), the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning, the Shuguang Program Project of Shanghai Educational Committee (Grant No. 14SG42).

References

[1] R. Albert, A.L. Barabási, Rev. Mod. Phys. 74 (2002) 47.

[2] M.E. Newman, SIAM Rev. 45 (2003) 167.
[3] J.G. Liu, Z.X. Wu, F. Wang, Int. J. Mod. Phys. C 18 (2007) 1087.
[4] D. Kempe, J. Kleinberg, É. Tardos, in: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2003, pp. 137–146.
[5] Z.L. Hu, J.G. Liu, G.Y. Yang, Z.M. Ren, Europhys. Lett. 106 (2014) 18002.
[6] M. Kitsak, L.K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H.E. Stanley, H.A. Makse, Nat. Phys. 6 (2010) 888.
[7] S. Pei, L. Muchnik, J.S. Andrade Jr., Z. Zheng, H.A. Makse, Sci. Rep. 4 (2014) 5547.
[8] Z.M. Ren, A. Zeng, D.B. Chen, H. Liao, J.G. Liu, Europhys. Lett. 106 (2014) 48005.
[9] J.G. Liu, Z.M. Ren, Q. Guo, Physica A 392 (2013) 4154.
[10] L.C. Freeman, Sociometry 40 (1977) 35.
[11] L.C. Freeman, D. Roeder, R.R. Mulholland, Soc. Netw. 2 (1980) 119.
[12] J. Wang, L. Rong, T. Guo, in: 4th International Conference on Wireless Communications, Networking and Mobile Computing, 2008, WiCOM'08, IEEE, 2008, pp. 1–4.
[13] J.H. Lin, Q. Guo, W.Z. Dong, L.Y. Tang, J.G. Liu, Phys. Lett. A 378 (2014) 3279.
[14] J.G. Liu, Z.M. Ren, Q. Guo, D.B. Chen, PLoS ONE 9 (2014) e104028.
[15] S. Pei, H.A. Makse, J. Stat. Mech. Theory Exp. 2013 (2013) P12002.
[16] Z.L. Hu, Z.M. Ren, G.Y. Yang, J.G. Liu, Int. J. Mod. Phys. C 25 (2014) 1440013.
[17] A. Rodriguez, A. Laio, Science 344 (2014) 1492.
[18] F. Morone, H.A. Makse, Nature 524 (2015) 65.
[19] X.Y. Zhao, B. Huang, M. Tang, H.F. Zhang, D.B. Chen, Europhys. Lett. 108 (2014) 68005.
[20] K. Appel, W. Haken, Ill. J. Math. 21 (1977) 429.
[21] G. Gonthier, Not. Am. Math. Soc. 55 (2008) 1382.
[22] B. Bollobás, Springer Science and Business Media, vol. 184, 1998, p. 57.
[23] D.J. Welsh, M.B. Powell, Comput. J. 10 (1967) 85.
[24] M.E. Newman, Phys. Rev. E 66 (2002) 016128.
[25] R.M. Anderson, R.M. May, B. Anderson, J. Public Health 16 (1992) 208.
[26] L.F. Zhong, J.G. Liu, M.S. Shang, Phys. Lett. A 379 (2015) 2272.
[27] Y. Liu, M. Tang, T. Zhou, Y. Do, Sci. Rep. 5 (2015) 13172.
[28] K. Klemm, M.Á. Serrano, V.M. Eguíluz, M. San Miguel, Sci. Rep. 2 (2012) 1.