LETTER

Identifying multiple influential spreaders via local structural similarity

To cite this article: J.-G. Liu *et al* 2017 *EPL* **119** 18001

A Letters Journal Exploring the Frontiers of Physics

July 2017

EPL, **119** (2017) 18001 www.epljournal.org
doi: 10.1209/0295-5075/119/18001

Identifying multiple influential spreaders via local structural similarity

J.-G. Liu[1(a)], Z.-Y. Wang[3], Q. Guo[2], L. Guo[2], Q. Chen[1] and Y.-Z. Ni[1]

[1] *Data Science and Cloud Service Centre, Shanghai University of Finance and Economics - Shanghai 200433, PRC*
[2] *Research Center of Complex Systems Science, University of Shanghai for Science and Technology Shanghai 200093, PRC*
[3] *College of Humanities, Shanghai University of Finance and Economics - Shanghai 200433, PRC*

received 28 April 2017; accepted in final form 9 August 2017
published online 7 September 2017

PACS 89.75.Fb – Structures and organization in complex systems
PACS 87.15.A- – Theory, modeling, and computer simulation

Abstract – Identifying the nodes with largest spreading influence is of significance for information diffusion, product exposure and contagious disease detection. In this letter, based on the local structural similarity, we present a method (LSS) to identify the multiple influential spreaders. Firstly, we choose the node with the largest degree as the first spreader. Then the new spreaders would be selected if they belong to the first- or second-order–neighbor node set of the spreaders and their local structural similarities with other spreaders are smaller than the threshold parameter r. Comparing with the susceptible-infected-recovered model, the experimental results for four empirical networks show that the spreading influences of spreaders selected by the local structural similarity method are larger than that of the color method, the degree, betweenness and closeness centralities. The further experimental results for the Barabàsi-Albert and random networks show that the LSS method could identify the multiple influential spreaders more accurately, which suggests that the local structural property plays a more important role than the distance for identifying multiple influential spreaders.

Introduction. – Identifying the influential spreaders is of significance for preventing epidemic spreading [1–3], controlling cascading failures in electrical power grids and Internet [4], which has attracted a great deal of attention in many fields [5–9]. So far, there are lots of methods presented to identify the influential spreaders in networks such as the degree [10,11], betweenness [12,13], eigenvector [14], closeness [15], k-shell decomposition [16–18], and so on [19–22]. Meanwhile, multiple influential spreader identification has attracted lots of attention. Hu *et al.* [23] identified the multiple influential spreaders in community networks and found that the distance among spreaders plays an important role. Rodriguez *et al.* [24] argued that large distance with higher densities could locate influential nodes and proposed a method to identify the influential spreaders. By taking the distance between spreaders into consideration, Zhao *et al.* [25] introduced the graph coloring to identify the influential spreaders. Guo *et al.* [26] proposed an improved distance-based coloring method to identify influential spreaders. Morone *et al.* [27] mapped the influence maximization into the optimal percolation problem to identify the minimal influential node set, which provides a theoretical framework to identify the influential spreaders. However, the distance calculation for larger-scale networks is very time-consuming and the threshold of the percolation method is hard to predefine. Besides the distance information, the local structural similarity also contains important information for measuring the relationship between each pair of spreaders, which can affect the performance of multiple spreaders.

Inspired by this idea, by introducing the local structural similarity, we present a method, namely the local structural similarity (LSS) method, to identify multiple influential spreaders. Specifically, we firstly choose a node with the largest degree as the first spreader. Then for the next target spreader who locates in the first- or second-order–neighbor node set of each existing spreader, we calculate its local structural similarities with each existing spreader. A node would be regarded as a new spreader only if its local structural similarities with all existing spreaders were

(a) E-mail: liujg004@ustc.edu.cn

smaller than the threshold parameter r. The experimental results of the LSS method for four empirical networks show that this method can identify the influential spreaders more accurately than the ones generated by the color method, degree, betweenness and closeness centralities.

The local structural similarity method. – Normally, an unweighted 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 [28]. Some scientists quantified a nodes' influence only accounting for its local surroundings [10,11], whose degree is introduced in this letter. In this letter, the LSS method is constructed as follows. Firstly, we choose a node with the largest degree to be the first spreader. Then for a target node belonging to the first- or second-order–neighbor node set of each existing spreader, we calculate its local structural similarity values with all existing spreaders. Finally, the target node would be chosen as a new spreader when its local structural similarity values with all existing spreaders are smaller than the threshold r. When the selection process stops, multiple influential spreaders can be obtained.

For a target node i, the local structural similarity value s_{ij} with the existing spreader j can be defined as

$$s_{ij} = \frac{|P_i \bigcap \Gamma_j|}{k_i}, \tag{1}$$

where $|\cdot|$ denotes number of nodes, k_i is the degree of node i, P_i denotes the nearest-neighbor node set of node i. When nodes i and j are connected, Γ_j denotes the first-order–neighbor node set of node j, while when node i is the second-order neighbor of node j, Γ_j denotes the first- and second-order neighbors of node j. Since the computation complexity of the LSS method would be very time-consuming when the distance between nodes i and j is larger than three, in this letter, we only take into account the case in which node i is one of the first- and second-order neighbors of node j.

The details of the presented LSS method are as follows:

Step 1: Ranking nodes in descending order in terms of the node degree, such that $k_1 \geq k_2 \geq \cdots \geq k_N$, where k_i denotes the degree of node i. The node at the top position of the degree ranking list is set as the first spreader,

Step 2: A new node i, locating in the first- or second-order–neighbor node set of all existing spreaders, is regarded as a new spreader if the local structural similarity s_{ij} between node i and all existing spreaders j is smaller than the threshold r;

Step 3: The chosen process will stop until there are no nodes locating in the first- or second-order–neighbor node set of the existing spreaders and and their local structural similarities are smaller than the threshold r.

Step 4: For all selected spreaders, the n nodes at the top of the list in terms of the selection process will be set as the initial spreaders.

From the local structure similarity s_{ij} and threshold parameter r, one can find that, when the structural similarity threshold is set as $r = 1$, the multiple spreaders chosen by the LSS method degenerate to the method in terms of the node degree. The closer to 0 the structural similarity threshold r is, the larger the structural differences among the spreaders are. The traditional methods are given as follows.

The coloring method. As known, for the graph coloring problem, 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 any two adjacent regions do not share the same color [29–31]. By taking the graph coloring problem into consideration, Zhao *et al.* [25] introduced the color method, namely the influential spreader (IS) method, to identify the influential spreaders. The main steps of the traditional IS method are as follows. Firstly, a network $G = (N, E)$ which could be described by an adjacent matrix $\mathbf{A} = a_{ij}$, where $a_{ij} = 1$ if node i is connected with node j, is colored with the rule that any two adjacent nodes do not share the same color [32]. When the coloring process ends, each node corresponds to a kind of color. Secondly, the nodes with the same color are classified as an independent set. Finally, the nodes at the top positions of the ranking list in the same color set are chosen as multiple spreaders.

The degree centrality. The node degree k_i is defined as the number of neighbors of node i, namely

$$k_i = \sum_{j=1}^{N} a_{ij}, \tag{2}$$

where a_{ij} is the element of matrix $\mathbf{A}$. Degree centrality can be a good measure in evaluating nodes' spreading influences. It is widely applied for its simplicity and low computational cost [13].

The betweenness centrality. The betweenness centrality measures the number of the shortest paths from a node to all others that pass through that node in a complex network [12], which could be denoted by

$$B_i = \sum_{s \neq i \neq t} \frac{n_{st}^i}{n_{st}}, \tag{3}$$

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

The closeness centrality. The closeness centrality of a node i is defined as the reciprocal of the sum of the shortest distances to all other nodes [15], which can be defined as

$$C_i = \sum_{j=1}^{N} \frac{N}{d_{ij}}, \tag{4}$$

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

Table 1: Basic statistical features of the Ca-GrQc, Routers, Soc-hamsterster and Polblogs networks, including the number of nodes N, the number of edges E, the average degree $\langle k\rangle = \frac{1}{N}\sum_i^N k_i$ and the average distance between each pair of nodes $\langle d\rangle$.

Network	N	E	$\langle k\rangle$	$\langle d\rangle$	$\beta_c = \frac{\langle k\rangle}{\langle k^2\rangle}$
Ca-GrQc	4158	13422	6.46	6.05	0.06
Routers	2113	6632	6.28	4.61	0.05
Soc-hamsterster	2000	16097	16.09	3.59	0.02
Polblogs	643	2280	7.09	3.83	0.04

The eigenvector. The eigenvalue was considered as the importance of a node is not only determined by itself, but it is also affected by its neighbors. Accordingly, the eigenvector centrality of node i, EC_i, is defined as

$$EC_i = \frac{1}{\lambda}\sum_{j=1}^{N} a_{ij}e_j, \tag{5}$$

where $EC = (e_1, e_2, \cdots, e_n)^T$, EC is the eigenvector of the adjacent matrix $\mathbf{A}$ corresponding to the largest eigenvalue λ.

The average shortest path length. The average shortest path length L_s among each pair of selected spreaders is also used to analyze the structural property of spreaders obtained by different methods, which is defined as

$$L_s = \frac{1}{|s|(|s|-1)}\sum_{\substack{u,v\in s\\ u\neq v}} d_{u,v}, \tag{6}$$

where s is the selected spreader set, $|s|$ denotes the number of spreaders in s and $d_{u,v}$ is the shortest distance between spreaders u and v.

In the following, we compare the performance of the LSS method and the traditional methods.

Experimental results. – Four empirical networks are introduced, including the Ca-GrQc, Routers, Soc-hamsterster and Polblogs networks. The Ca-GrQc network is a collaboration network in which node and link represent the individual and scientific collaboration. The Routers network is a technological network where the link represents the communication between different routers. The Soc-hamsterster network is a social network that the link represents the friendship and family link. And the Polblogs network is a communication relationship network. The node represents the owner blog and the link represents the communication. The four data sets can be accessible from the web site `http://networkrepository.com/`. The statistical properties of the four empirical networks are shown in table 1.

The spreading model. We employ the susceptible-infected-recovered (SIR) [33] model to simulate the spreading process. In the SIR model, the nodes have three states [34]: i) susceptible individuals represent the individuals (not yet infected) who are likely 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, we denote that all nodes are initially susceptible only except for the initial nodes. And then 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 probability μ. Finally, if the spreading process reaches the steady state, the number of nodes generated by the initially infected node including the initially infected node is defined as the spreading influence of the target node, which is denoted by R_{SI}.

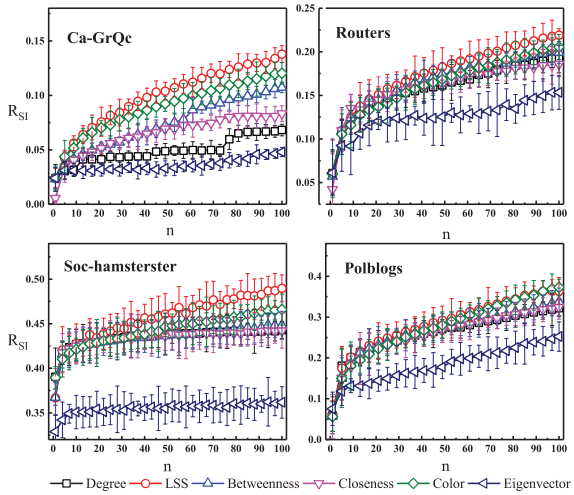

Fig. 1: (Color online) The spreading influence R_{SI} of spreaders to different spreaders n for four networks at the optimal threshold r, where the optimal structural similarity threshold r is 0.6, 0.6, 0.7, 0.8 for the Ca-GrQc, Routers, Soc-hamsterster and Polblogs networks, respectively. The results are averaged over 1000 independent runs with the spreading rate $\beta = 0.01$ and $\mu = 0.1$.

Results and analysis. We compare the performance of the LSS method in the SIR model with the performance of the color method, the degree, betweenness and closeness centralities for four empirical networks. Figure 1 shows the spreading influence R_{SI} of multiple spreaders for different numbers of spreaders n. One can find that for Ca-GrQc, Routers, Soc-hamsterster and Polblogs networks, the spreading influence R_{SI} of the LSS method is larger than the ones obtained by the color method, the degree, betweenness and closeness centralities, indicating that the LSS method performs better than other methods. Specifically, for the Ca-GrQc network, when the number of spreaders $n = 100$ and the effective transmission rate $\lambda = 0.1$, the spreading influence R_{SI} of the LSS method can reach 0.138, while the ones of the color

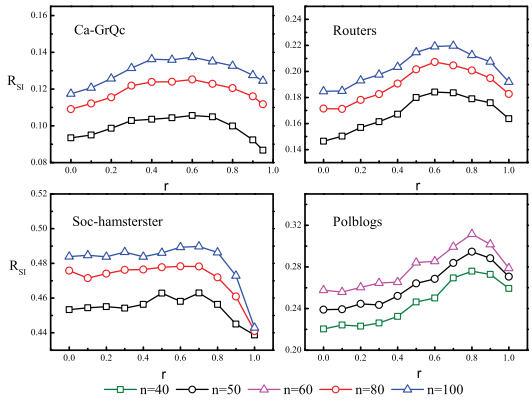

Fig. 2: (Color online) The spreading influence R_{SI} of spreaders for different threshold parameter r. The parameter n is set as the number of spreaders which is selected from the spreader list accordingly. The results are averaged over 1000 independent runs with the spreading rate $\beta = 0.01$ and $\mu = 0.1$.

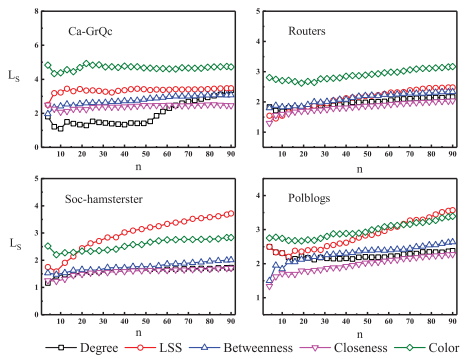

Fig. 3: (Color online) The average shortest path length L_s among spreaders obtained by the LSS method, color method, the degree, betweenness and closeness centralities.

method, degree, betweenness and closeness centralities are 0.120, 0.068, 0.107, 0.083. The same results can be obtained from fig. 1 for Routers, Soc-hamsterster and Polblogs networks. For larger spreading rate β, the spreaders generated by the LSS method have larger influence than the ones obtained by other methods.

Furthermore, we analyse the impact of the structural similarity threshold r on the performance of the LSS method in the SIR model. From fig. 2, one can find that the spreading influence of the LSS method increases to the largest value with the structural similarity threshold r and then decreases. There is an optimal structural similarity threshold r to the maximum the spreading influence R_{SI}. Specifically, the optimal structural similarity threshold r is 0.6, 0.6, 0.7, 0.8 for Ca-GrQc, Routers, Soc-hamsterster and Polblogs networks, respectively, which indicates that there is an optimal structural similarity threshold r for different networks.

Figure 3 shows the average shortest path length L_s between the spreaders obtained by different methods, from which one can find that, for Ca-GrQc, Polblogs and Routers networks, the distances between each pair of spreaders obtained by the LSS method are larger than the ones of degree, betweenness and closeness centralities, but smaller than the ones obtained by the color method. While for the Soc-hamsterster network, the distances between each pair of spreaders gotten by the LSS method are larger than the ones of other methods, which indicates that the structural similarity among spreaders plays a more important role than the distance for choosing multiple influential spreaders.

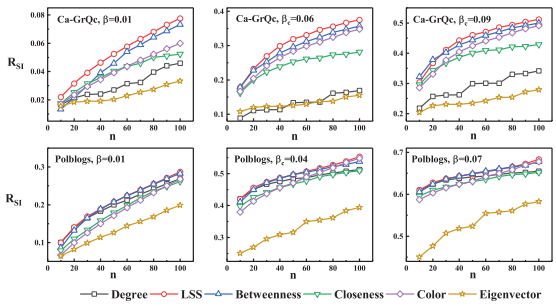

Fig. 4: (Color online) The spreading influence R_{SI} of spreaders for different spreading rates $\beta = 0.01$, $\beta_c = 0.04$ and $\beta = 0.07$ for the Polblogs data set, from which one can find that the spreading influence R_{SI} outperforms other indices for different parameters n and β.

The analyses for the four empirical networks show that the LSS method performs better than the color method, the degree, betweenness and closeness centralities. In addition, there is an optimal structural similarity threshold r where the spreading influence R_{SI} can reach its maximum value. The reason why the LSS method performs better maybe lies in the fact that the color method fixes the distances between each pair of spreaders as a constant value while the LSS method chooses influential nodes depending on the local structural similarity with spreaders. When two influential nodes are very close, one of these two nodes cannot be chosen as a spreader by the color method, while two influential nodes can be regarded as spreaders simultaneously by considering local structural similarity with spreaders.

Figure 4 presents the spreading influence R_{SI} for different spreading rates β, from which one can find that the LSS method outperforms the color method, degree, betweenness and eigenvector indices for different parameters β. Take the Polblogs data set as an example, the spreaders selected according to the LSS could affect more nodes for different parameters β, including the small spreading rate $\beta = 0.01$, the threshold $\beta_c = 0.04$ and the larger spreading rate $\beta = 0.07$.

Furthermore, we investigate the performance of the LSS method for the Erdös-Rényi (ER) network [35] and the Barabási-Albert (BA) networks [7], where the ER network is constructed with the connect possibility $p = 0.01$ and networks size $N = 300$, and the BA network is constructed with the $m = 3$ and network size $N = 500$.

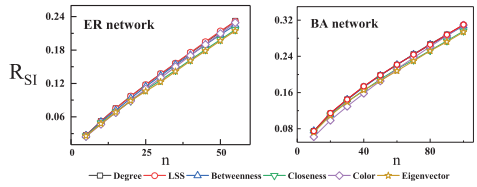

Fig. 5: (Color online) The spreading influence R_{SI} of spreaders for different numbers of spreaders n for the Erdös-Rényi random network and the Barabási-Albert scale-free networks when $\beta = 0.01$ and $\mu = 0.1$. One can find that, although the performances are very close, comparing with the degree, betweenness, eigenvalue and color methods, the spreaders identified by the LSS method still have larger spreading influence R_{SI}.

The comparisons among the LSS and other methods are given in fig. 5, from which one can find that, for the synthetic ER and BA networks, the presented LSS method still has larger spreading influence then other methods.

Conclusion and discussions. – Identifying multiple influential spreaders in networks is an important task to control the diffusion process. In this letter, by taking into account the local structural similarity between each pair of spreaders, we propose the LSS method to identify multiple influential spreaders. A target node would be regarded as a spreader if it belongs to the first or second-order–neighbor node set of other spreaders and its local similarities value with other spreaders are smaller than the threshold r. Comparing with the color method, the degree, betweenness and closeness centralities, the experimental results for four empirical networks in the SIR and SI models show that the performance of the LSS method performs better than the ones of other methods. And for different networks, each has an optimal structural similarity threshold parameter whose spreading influence can reach the maximum value. In order to analyze the structural property of influential nodes, the average shortest path length is introduced. The empirical results show that the distance among spreaders obtained by the LSS method is larger than that of other methods, which suggests that the local similarity could effectively identify the multiple spreaders. Furthermore, the effect of the parameter β for the spreading influence R_{SI} is investigated. The experimental results for the Polblogs data set show that the performance of the LSS method outperforms the degree, betweenness, eigenvector and color methods when parameter β is smaller or larger than the threshold β_c. Then we construct an Erdös-Rényi random network and Barabási-Albert scale-free networks. The experimental results for the random and scale-free networks also show that the LSS method could identify the multiple spreaders more accurately.

The node spreading influence is determined not only by the network structure, but also by the spreading dynamics [36,37]. Therefore, in terms of the local structure information, spreading dynamics is a promising way for identifying the multiple spreaders. The results of the LSS method depend on the first-spreader selection. How to select the first spreader to increase the performance is an open question. For small threshold r, the distance between each pair of spreaders would increase, while for large threshold r, the spreaders would be too close. So it would be a challenge work to analyze the relationship between the distance among spreaders and the threshold r. There are other properties of a network that may affect the performance of the LSS method. For example, the sparsity and density of a network may affect the performance of the LSS method. Taking into account the node mobility [38], how to effectively identify the influential spreaders is another important question. The future work would also focus on how to analyze the relationship between the network structure and the optimal distance between multiple spreaders.

* * *

This work is supported by the National Natural Science Foundation of China (Grant Nos. 61374177, 71371125), 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). J-GL is supported by the funding of SHUFE (Grant No. 2017110022).

REFERENCES

[1] Paster-Satorras R., Castellano C., Mieghem Ven P. and Vespignani A., *Rev. Mod. Phys.*, **87** (2015) 925.
[2] Liu J. G., Lin J. H., Guo Q. and Zhou T., *Sci. Rep.*, **6** (2016) 21380.
[3] Kempe D., Kleinberg J. and Tardos É., *Maximizing the Spread of Influence through a Social Network*, in *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York) 2003, p. 137.
[4] Motter A. E., *Phys. Rev. Lett.*, **93** (2004) 098701.
[5] Zhou T., Liu J. G., Bai W. J., Chen G. and Wang B. H., *Phys. Rev. E*, **74** (2006) 056109.
[6] Siganos G., Faloutsos M., Faloutsos P. and Faloutsos C., *IEEE/ACM Trans. Netw.*, **11** (2002) 514.
[7] Albert R. and Barabási A. L., *Rev. Mod. Phys.*, **74** (2002) 47.
[8] Du Y., Gao C., Hu Y., Mahadevan S. and Deng Y., *Physica A*, **339** (2014) 57.
[9] Zhou T., Fu Z. H. and Wang B. H., *Prog. Nat. Sci.*, **16** (2006) 452.
[10] Newman M. E., *SIAM Rev.*, **45** (2003) 167.
[11] Pei S., Muchnik L., Andrade J. S. jr., Zheng Z. and Makse H. A., *Sci. Rep.*, **4** (2014) 5547.
[12] Freeman L. C., *Sociometry*, **40** (1977) 35.
[13] Wang J., Rong L. and Guo T., *A new measure of node importance in complex networks with tunable parameters*, in *4th International Conference on Wireless Communications, Networking and Mobile Computing, 2008. WiCOM'08* (IEEE) 2008, pp. 1–4.

[14] Estrada E. and Rodríguez-Velázquez J. A., *Phys. Rev. E*, **71** (2005) 056103.
[15] Freeman L. T. C., Roeder D. and Mulholland R. R., *Soc. Net.*, **2** (1979) 119.
[16] Kitsak M., Gallos L. K., Havlin S., Liljeros F., Muchnik L., Stanley H. E. and Makse H. A., *Nat. Phys.*, **6** (2010) 888.
[17] Lin J. H., Guo Q., Dong W. Z., Tang L. Y. and Liu J. G., *Phys. Lett. A*, **378** (2014) 3279.
[18] Liu J. G., Ren Z. M., Guo Q. and Chen D. B., *PLoS ONE*, **9** (2014) e104028.
[19] Lv L., Zhou T., Zhang Q. M. and Stanley H. E., *Nat. Commun.*, **7** (2016) 10168.
[20] Malliaros F. D., Rossi M. E. G. and Vazirgiannis M., *Sci. Rep.*, **6** (2016) 19307.
[21] Shuang X., Wang P. and Lv J. H., *Sci. Rep.*, **7** (2017) 41321.
[22] Zhang J. X., Chen D. B., Dong Q. and Zhao Z. D., *Sci. Rep.*, **6** (2016) 27823.
[23] Hu Z. L., Liu J. G., Yang G. Y. and Ren Z. M., *EPL*, **106** (2014) 18002.
[24] Rodriguez A. and Laio A., *Science*, **344** (2014) 1492.
[25] Zhao X. Y., Huang B., Tang M., Zhang H. F. and Chen D. B., *EPL*, **108** (2014) 68005.
[26] Guo L., Lin J. H., Guo Q. and Liu J. G., *Phys. Lett. A*, **380** (2015) 837.
[27] Morone F. and Makse H. A., *Nature*, **542** (2015) 65.
[28] Ren Z. M., Zeng A., Chen D. B., Liao H. and Liu J. G., *EPL*, **106** (2014) 48005.
[29] Appel K. and Haken W., *Ill. J. Math.*, **21** (1977) 429.
[30] Gonthier G., *Not. AMS*, **55** (2008) 1382.
[31] Bollobás B., *Springer Sci. Bus. Media*, **184** (1998) 57.
[32] Welsh D. J. and Powell M. B., *Comput. J.*, **10** (1967) 85.
[33] Newman M. E., *Phys. Rev. E*, **66** (2002) 016128.
[34] Anderson R. M., May R. M. and Anderson B., *Aus. J. Public Health*, **16** (1992) 208.
[35] Erdös P. and Rényi A., *Math. Debrecen*, **6** (1959) 290.
[36] Zhong L. F., Liu J. G. and Shang M. S., *Phys. Lett. A*, **379** (2015) 2272.
[37] Liu Y., Tang M., Zhou T. and Do Y., *Sci. Rep.*, **5** (2015) 13172.
[38] Liu J. G., Wu Z. X. and Wang F., *Int. J. Mod. Phys. C*, **18** (2007) 1087.