Ranking the spreading influence in complex networks

Jian-Guo Liu *, Zhuo-Ming Ren, Qiang Guo

Research Center of Complex Systems Science, University of Shanghai for Science and Technology, Shanghai 200093, PR China

HIGHLIGHTS

- K-shell decomposition method could only identify the network-core successfully.
- The spreading influence of the nodes with same k-shell values are very different.
- We present a ranking way to identify the spreading influence of the nodes with same k-shell values.
- The proposed method could rank the node spreading influence more accurately than the k-shell, degree k and other methods.

ARTICLE INFO

ABSTRACT

Identifying the node spreading influence in networks is an important task to optimally use the network structure and ensure the more efficient spreading in information. In this paper, by taking into account the shortest distance between a target node and the node set with the highest k-core value, we present an improved method to generate the ranking list to evaluate the node spreading influence. Comparing with the epidemic process results for four real networks and the Barabási–Albert network, the parameterless method could identify the node spreading influence more accurately than the ones generated by the degree k, closeness centrality, k-shell and mixed degree decomposition methods. This work would be helpful for deeply understanding the node importance of a network.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

Spreading is an ubiquitous process in nature, which describes many important activities in society [1–4], such as the virus spreading [5,6], reaction diffusion processes [7,8], pandemics [9], cascading failures [10] and so on. The knowledge of the spreading pathways through the network of interactions is crucial for developing effective methods to either hinder the disease spreading, or accelerate the information dissemination spreading. So far, a lot of works focus on identifying the most influential spreaders in a network [11–13], for example, the most connected nodes (hubs) are supposed to be the key spreaders, being responsible for the largest scale of the spreading process [14–16]. Recently, Kitsak et al. [1] argued that the node spreading influence is determined by its location in a network. By decomposing a network with the k-shell decomposition method, they found that the most influential nodes, namely the network core, could be identified by the largest k-core values. It should be noticed that the k-shell method assigns many nodes with the same k-core value even though they perform entirely differently in the spreading process. Fig. 1 shows that, for some real networks, there are lots of nodes whose k-core values, denoted as k_s, are equal. By taking into the number of removed and existed links in the decomposition process, Zeng et al. [17] proposed an improved method, named mixed degree decomposition (MDD) method, to distinguish the node spreading influence within the node set with the same k_s value. To different networks, the optimal parameters of the MDD method are determined by the statistical properties of the networks, which hinder its application. By investigating the effects of privileged spreaders on social networks, Borge et al. [12,13] found that the node spreading

---

* Corresponding author. Tel.: +86 15202121269.
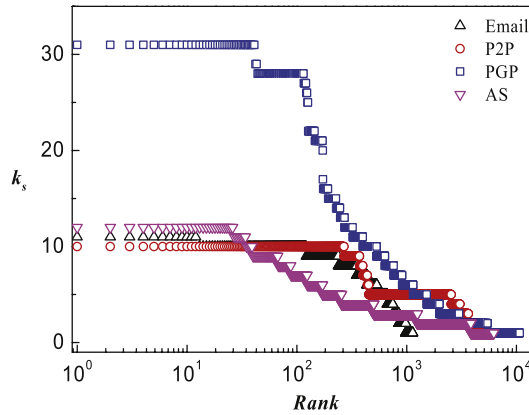E-mail address: liujg004@ustc.edu.cn (J.-G. Liu).

Fig. 1. (Color online) Rank of the k_s values for Email, P2P, PGP, and AS networks, from which one can find that there are lots of nodes whose k_s values are equal.

influence does not depend on their k-core values, which instead determines whether or not a given node prevents the diffusion process. These literatures suggest that, besides the network core, it is also very important to generate a ranking list to identify all nodes' spreading influences. In this paper, we argue that, for the node set with same k-core values, the nodes whose locations are close to the network core have larger spreading influences. Inspired by the idea, we present an improved k-shell method to generate the global influential ranking list. Comparing with the susceptible-infection-recovered (SIR) spreading process [18,19] for four real networks and the Barabási–Albert network [20], the experimental results show that our method could generate the ranking list more accurately than the ones generated by the degree k, closeness centrality(CC) [21], k-shell and MDD decomposition methods respectively.

2. Method

Normally, a network $G = (N, E)$ with N nodes and E links could be described by an adjacent matrix $A = \{a_{ij}\} \in R^{n,n}$, where $a_{ij} = 1$ if node i is connected by node j, and $a_{ij} = 0$ otherwise. The node degree k_i is defined as the number of neighbors for node i. The closeness centrality(CC) of node i is defined as the reciprocal of the sum of the shortest distances to all other nodes of N [21]. The k-shell decomposition method [22,23] could be implemented in the following way to identify the network core. Firstly, remove all nodes with degree one, and then keep pruning the existed nodes until all nodes' degrees are larger than one. The removed nodes would form a node set whose k-core value equals to one. Then, repeat the pruning process in the same way for the rest nodes. Finally, the k-shell method decomposes a network into different node set with different k-core values. Implementing the SIR spreading process for one network, one can find that the nodes with the same k_s values always have different number of infected nodes, namely *spreading influence*. This phenomena suggests that the k-shell decomposition method is not appropriate for ranking the global spreading influence of a network. In terms of the distance from a target node to the network core, the spreading influences of the nodes with the same k-core values could be distinguished in the following way

$$\theta(i|k_s) = (k_s^{\max} - k_s + 1) \sum_{j \in J} d_{ij}, \quad i \in S_{k_s} \tag{1}$$

where $k_s^{\max}$ is the largest k-core value of a network, the shortest distance d_{ij} is measured by the shortest distance from the node i to the node j, J is denoted as the network core node set, and S_{k_s} is denoted as the node set whose k-core values equal to k_s.

3. How to evaluate the performance

To check the performance of the improved method, four real networks are introduced in this paper, which include (i) *Email network* [24]. The Email network of University Rovira i Virgili (URV) of Spain contains faculty, researchers, technicians, managers, administrators, and graduate students. (ii) *Peer-To-Peer (P2P) network* [25]. A sequence of snapshots of the Gnutella peer-to-peer file sharing network from August 2002. Each node represents a host in the Gnutella network and each link represents the connection between each pair of Gnutella hosts. (iii) *Pretty-Good-Privacy (PGP) network* [26]. Pretty-Good-Privacy algorithms have been developed in order to maintain privacy between peers, wherefore, it is also called web of trust of PGP. (iv)*Autonomous Systems (AS) network* [27]. The data was collected from University of Oregon Route Views Project — Online data and reports. The network of routers comprising the Internet can be organized into sub-graphs called Autonomous Systems. Each AS exchanges traffic flows with some neighbors. The statistical properties include the number of nodes N and links E of the network, the average degree, the second-order average degree, and the spreading threshold are given in Table 1.

Table 1
The statistical properties of the networks, where N is the number of nodes, E is the number of links, $\langle k \rangle = \frac{1}{N}\sum_i k_i$ and $\langle k^2 \rangle = \frac{1}{N}\sum_i k_i^2$ are the average degree and the second-order average degree of the network, β_{rand}^c is the spreading threshold for a network [28] ($\beta_{\text{rand}}^c \approx \langle k \rangle / \langle k^2 \rangle$).

Networks	N	E	$\langle k \rangle$	$\langle k^2 \rangle$	β_{rand}^c
Email	1 133	5 451	9.60	180	0.053
P2P	6 301	20 778	7.00	116	0.060
PGP	10 680	24 316	4.60	86	0.053
AS	6 202	12 170	3.92	618	0.006

To evaluate the performance of the improved method, the Kendall's tau [29,30] is introduced to measure the accuracy of the method. By using the degree k, k-shell and MDD methods, we could obtain different ranking lists in terms of the network structure. In principle, the ranking lists generated by an effective structure-based ranking method should be as close as possible to the ranking list generated by the real spreading process. In this paper, we employ the SIR model [1] to simulate the spreading process on networks. In the SIR model, we denote that all nodes are initially susceptible except only one infectious node i. In each time step, the infected nodes will infect their susceptible neighbors with the spreading rate β, and infected nodes would recover in two time steps [1]. The number of infections s_i^β generated by the initially-infected node i is denoted as its spreading influence, where β is the spreading rate in the SIR model. Ranking the node spreading influence in terms of its spreading influence s_i^β, one could obtain the ranking list of the SIR-model-based spreading influence. We therefore use the Kendall's tau coefficient τ to measure the correlation between one topology-based ranking list and the one generated by the SIR model. The higher the Kendall's tau value τ is, the more accurate result the method could generate. The most ideal case, $\tau = 1$, indicates that the method uniquely identifies the real influence ranking list.

4. Numerical results

For Email, P2P, PGP, and AS networks, the Kendall's tau values τ for the degree k, CC, k-core and MDD indices are shown in Fig. 2, from which one can find that, when the spreading rate β is higher than the epidemic threshold β_{rand}^c (the dot line), the Kendall's tau value τ of the improved method θ would be much better than the other indices. When the spreading rate β is much smaller than the epidemic threshold β_{rand}, the SIR process would stop in a few first infection steps, therefore the node with large degree k would infect more nodes, which may be the reason why the τ of the degree k is very large when the spreading rate β is much smaller than β_{rand}^c. The comparisons between the SIR model and the improved method show that the nodes who are closer to the network core have a larger spreading influence, which is consistent with the main idea.

Besides the real networks, we also investigate the performance of the improved method θ for the Barabási–Albert network [20,31], namely BA network. In the BA network, there are m_0 nodes in initial condition. In each time step, a new node with m links would connect the existing nodes according to the preferential attachment mechanism. By using the k-shell decomposition method for a BA network, one can find that all nodes have the same $k_s = m$ values except the initial m_0 nodes, which indicates that the traditional k-shell method could not be used to analyze this kind of networks. By implementing our improved method, one can find that the node spreading influence could be ranked more accurately than the degree k and MDD methods. Fig. 3 shows that the Kendall's tau values τ generated by the degree k and MDD methods are equal for different spreading rate β for the BA networks, which indicates that the MDD method would degenerate to the degree k index for the BA network. One also could find that, when the spreading rate β is larger than the spreading threshold β_{rand}^c, the Kendall's tau values τ would increase accordingly. It should be emphasized that when β is very large, there would be a crossover for the Kendall's tau values generated by the θ and the ones generated by the degree k, MDD methods. The reason may lie in the fact that when the spreading rate β is very large, the network core could very easily be infected by the nodes with largest degrees, and then the infections would spread to the entire network very quickly.

In order to investigate the capability of the method to distinguish the spreading influences of the nodes with same k-core values, we define the *distinct* metric D as follows

$$D = \sum_{k_s^{\min}}^{k_s^{\max}} \frac{\text{\# distinct elements in } S_{k_s}}{N} \tag{2}$$

where $k_s^{\min}$ and $k_s^{\max}$ denote the minimum and maximum k-core values of a network. For the nodes belongs to the node set S_{k_s}, the number of distinct elements could be used to measure the improvement of the method to the traditional k-shell decomposition method. For example, for a node set $\{1, 2, 3\} \in S_1$ whose degrees are 5,2,5, one can find that there are two distinct degree values $\{2, 5\}$, therefore, the number of distinct elements for S_1 equals to two. The largest value $D = 1$ indicates that all nodes of each node set S_{k_s} ($k_s \in [k_s^{\min}, k_s^{\max}]$) could be identically distinguished, while the minimum value $D = \frac{1}{N}$ means that all nodes are assigned the same k_s value.

Table 2 shows the results of the distinct D for four real networks, which indicate that, comparing with the degree k and MDD methods, our improved method θ could be much better at distinguishing the node spreading influences for the nodes with the same k_s value.
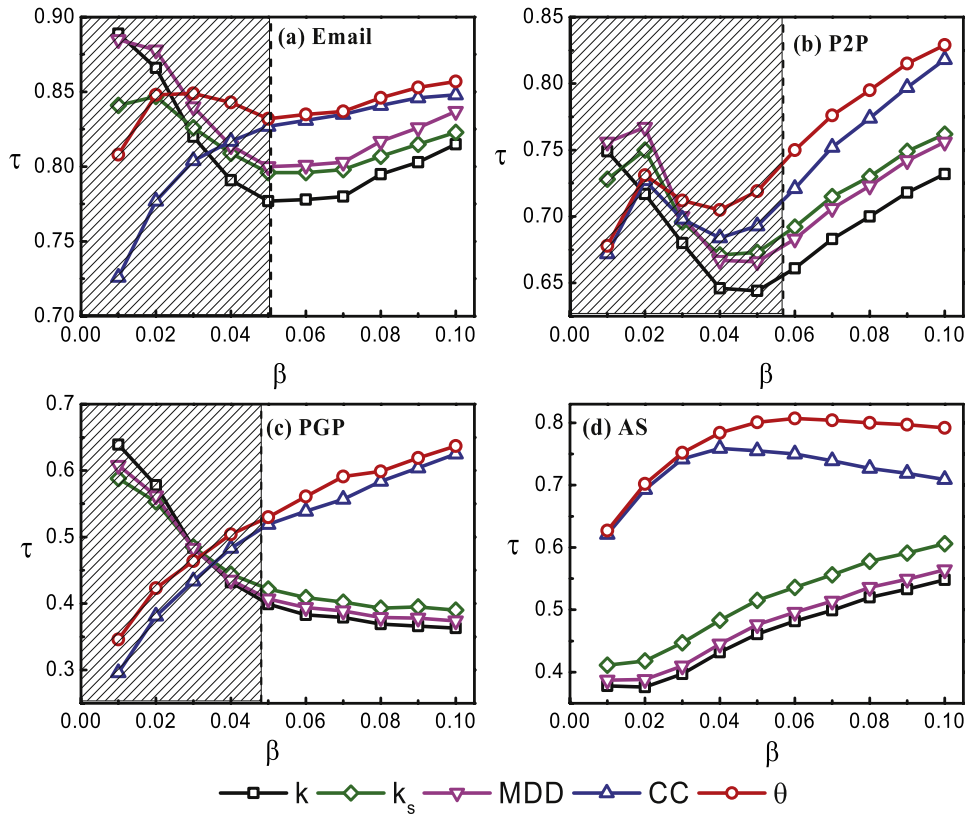
Fig. 2. (Color online) The Kendall's tau values τ obtained by comparing the ranking list generated by the SIR spreading process and the ranking lists generated by the improved method θ for Email, P2P, PGP, and AS networks, where the dot line corresponds to the threshold β_{rand}^c. From which one can find that the new method could identify the node spreading influence more accurately when the spreading rate β is larger than the threshold β_{rand}^c. The results are averaged over 100 independent runs with different spreading rate β.

Table 2
The distinct D values for Email, P2P, PGP and AS networks.

Index	Email (%)	P2P (%)	PGP (%)	AS (%)
k	12.27	2.57	3.76	3.58
MDD	20.92	5.33	3.57	2.95
θ	**42.81**	**28.92**	**13.55**	**4.60**

5. Conclusions an discussions

In summary, we propose a parameterless method to rank the node spreading influence in terms of the node distance to the network core which is defined as the node set with highest k_s values. The k-shell decomposition method could identify the most influential spreaders of a network, and also assign some nodes with the same value regardless their characters in the spreading process. According to the SIR spreading process results, one can find that the nodes with the same k_s value have very different spreading influences. The nodes whose locations are close to the network core play more significant role in the spreading process. Taking into account the k_s values and the shortest distance between the target node and the network core, we propose an improved method to rank the node spreading influence. The simulation results for four real networks and the BA network show that, comparing with the SIR spreading process results, this method could identify the node spreading influence more accurately than the degree k, closeness centrality, k-shell and MDD decomposition methods. Our method is parameterless and only depends on the k_s value and the distance from the target node to the network core, which is very helpful for the widely application in real systems.

Being great value to practice and theory, several methods are proposed to rank the node spreading influence. However, the presented methods so far mainly focus on the node degrees or positions. We here turn to a new perspective to understand the relationship between not only the k-shell location, but the nodes' shortest distance to the network core. Up to now, although the distance-based method could shed some light on how the position and distance to the network core affect the node spreading influence, we still lack systematic comparison and understanding of the performances of these measures, which is set as our future work. Klemm et al. [2] argued that the importance of a node in a network is not uniquely determined
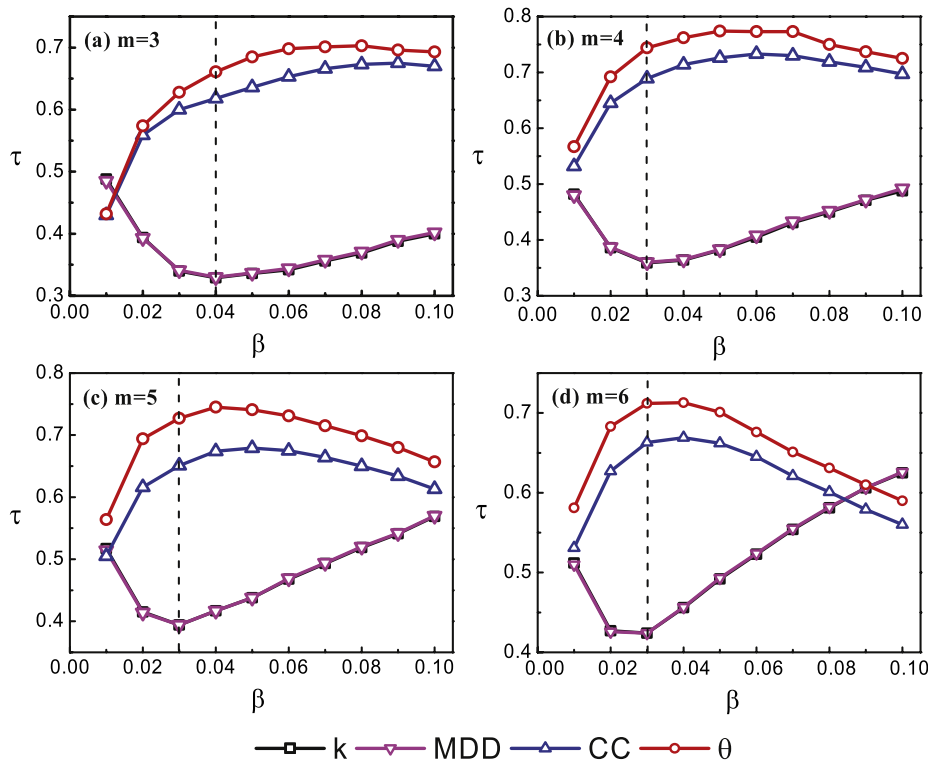
Fig. 3. (Color online) Comparing the influential ranking lists generated by the improved method and the SIR result, the Kendall's tau values τ for different BA networks, where the dot line corresponds to the threshold β_{rand}^c. The parameters are set as $N = 7000$, $m_0 = 20$ and $k_s = m = 3, 4, 5, 6, 7$. From which one can find that the τ values of the degree and MDD methods are equal, which indicates that the MDD method degenerates to the degree k for the BA networks, and the τ values of the improved method are much larger than the degree k and MDD methods when the spreading rate β is larger than the threshold β_{rand}^c and smaller than the crossover. The results are averaged over 100 independent runs with different spreading rates.

by the system structure, but it is a result of the interplay between dynamics and network structure. Empirical analysis on more known and proposed indices as well as more dynamic models [32,33] is very valuable for deeply understanding the spreading dynamic and building up knowledge and experience. A clear picture of this issue can be completed by putting together many fragments from respective empirical studies. Besides the empirical results, an alternative way is to build artificial network models with controllable topological features. In this way, we could have a clear picture on the unknown and uncontrollable ingredients which are always mixed together in real networks.

Acknowledgments

References

[1] M. Kitsak, L.K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H.E. Stanley, H.A. Makse, Identification of influential spreaders in complex networks, Nature Physics 6 (11) (2010) 888–893.
[2] K. Klemm, M. Serrano, V. Eguiluz, M. Miguel, A measure of individual role in collective dynamics, Scientific Reports 2 (2012) 292.
[3] C. Castellano, R. Pastor-Satorras, Competing activation mechanisms in epidemics on networks, Scientific Reports 2 (2012) 371.
[4] T. Zhou, J.-G. Liu, W.-J. Bai, G. Chen, B.-H. Wang, Behaviors of susceptible-infected epidemics on scale-free networks with identical infectivity, Physical Review E 74 (5) (2006) 056109.
[5] F.B. Cohen, A Short Course on Computer Viruses, John Wiley & Sons, New York, 1994.
[6] J.O. Kephart, G.B. Sorkin, D.M. Chess, S.R. White, Fighting computer viruses, Scientific American 277 (5) (1997) 56–61.
[7] V. Colizza, R. Pastor-Satorras, A. Vespignani, Reaction diffusion processes and metapopulation models in heterogeneous networks, Nature Physics 3 (4) (2007) 276–282.
[8] J.-G. Liu, Z.-X. Wu, F. Wang, Opinion spreading and consensus formation on square lattice, International Journal of Modern Physics C 18 (07) (2007) 1087–1094.
[9] R. Pastor-Satorras, A. Vespignani, Dynamical and correlation properties of the internet, Physical Review Letters 87 (25) (2001) 258701.
[10] A.E. Motter, Cascade control and defense in complex networks, Physical Review Letters 93 (9) (2004) 098701.
[11] G. Ghoshal, A.L. Barabási, Ranking stability and super-stable nodes in complex networks, Nature Communications 2 (2011) 394.
[12] J. Borge-Holthoefer, Y. Moreno, Absence of influential spreaders in rumor dynamics, Physical Review E 85 (2) (2012) 026116.

[13] J. Borge-Holthoefer, A. Rivero, Y. Moreno, Locating privileged spreaders on an online social network, Physical Review E 85 (6) (2012) 066123.
[14] R. Albert, H. Jeong, A.L. Barabási, Error and attack tolerance of complex networks, Nature 406 (6794) (2000) 378–382.
[15] R. Cohen, K. Erez, D. Ben-Avraham, S. Havlin, Breakdown of the internet under intentional attack, Physical Review Letters 86 (16) (2001) 3682–3685.
[16] J.P. Gleeson, D.J. Cahalane, Seed size strongly affects cascades on random networks, Physical Review E 75 (5) (2007) 056103.
[17] A. Zeng, C.J. Zhang, Ranking spreaders by decomposing complex networks, Physics Letters A 377 (14) (2013) 1031–1035.
[18] A. Barrat, M. Barthlemy, A. Vespignani, Dynamical Processes on Complex Networks, Cambridge Univ., 2008.
[19] M.E.J. Newman, Spread of epidemic disease on networks, Physical Review E 66 (1) (2002) 016128.
[20] A.L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (5439) (1999) 509–512.
[21] G. Sabidussi, The centrality index of a graph, Psychometrika 31 (4) (1966) 581–603.
[22] S.N. Dorogovtsev, A.V. Goltsev, J.F.F. Mendes, K-core organization of complex networks, Physical Review Letters 96 (4) (2006) 040601.
[23] S. Carmi, S. Havlin, S. Kirkpatrick, S.Y. Havitt, E. Shir, A model of Internet topology using k-shell decomposition, Proceedings of the National Academy of Sciences 104 (27) (2007) 11150–11154.
[24] R. Guimera, L. Danon, A. Diaz-Guilera, F. Giralt, A. Arenas, Self-similar community structure in a network of human interactions, Physical Review E 68 (6) (2003) 065103.
[25] J. Leskovec, J. Kleinberg, C. Faloutsos, Graph evolution: densification and shrinking diameters, ACM Transactions on Knowledge Discovery from Data (TKDD) 1 (1) (2007) 2.
[26] M. Boguna, R. Pastor-Satorras, A. Daz-Guilera, A. Arenas, Models of social networks based on social distance attachment, Physical Review E 70 (5) (2004) 056122.
[27] J. Leskovec, J. Kleinberg, C. Faloutsos, Graphs over time: densification laws, shrinking diameters and possible explanations, in: Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, ACM, 2005, pp. 177–187.
[28] C. Castellano, R. Pastor-Satorras, Thresholds for epidemic spreading in networks, Physical Review Letters 105 (21) (2010) 218701.
[29] R.B. Nelsen, Kendall Tau Metric, Springer, Michiel, 2001.
[30] Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be a set of joint observations from two random variables X and Y respectively, If one has $x_i > x_j$ and $y_i > y_j$ or $x_i < x_j$ and $y_i < y_j$, the observations (x_i, y_i) and (x_j, y_j) are said to be concordant. If $x_i > x_j$ and $y_i < y_j$ or if both $x_i < x_j$ and $y_i > y_j$, the ranks for both elements are said to be discordant. If $x_i = x_j$ and $y_i = y_j$, the pair is neither concordant nor discordant. Then Kendall's tau could be defined as $\frac{N_c - N_d}{\sqrt{(n_0 - n_1)(n_0 - n_2)}}$, where $n_0 = n(n-1)/2$, $n_1 = \sum_i t_i(t_i - 1)/2$, $n_2 = \sum_j t_j(t_j - 1)/2$, N_c and N_d are the number of concordant and discordant pairs, t_i and t_j are the number of tied values in the ith and jth group of ties respectively.
[31] R. Albert, A.L. Barabási, Statistical mechanics of complex networks, Reviews of Modern Physics 74 (1) (2002) 47.
[32] Y.C. Zhang, M. Blattner, Y.K. Yu, Heat conduction process on community networks as a recommendation model, Physical Review Letters 99 (15) (2007) 154301.
[33] J.-G. Liu, T. Zhou, Q. Guo, Information filtering via biased heat conduction, Physical Review E 84 (3) (2011) 037101.