CHIN. PHYS. LETT. Vol. 33, No. 2 (2016) 028901

# Identifying Influence of Nodes in Complex Networks with Coreness Centrality: Decreasing the Impact of Densely Local Connection

Yi-Run Ruan(阮逸润)**, Song-Yang Lao(老松杨), Yan-Dong Xiao(肖延东), Jun-De Wang(王竣德), Liang Bai(白亮)

*Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, Changsha 410073*

(Received 28 September 2015)

*Ranking nodes by their spreading ability has been a fundamental issue related to many real applications, such as information and disease control. In the well-known coreness centrality measure, nodes' influence is ranked only by summing all neighbors' k-shell values while the effect of the topological connections among them on the nodes' spreading ability are ignored. In this work, we propose an improved coreness measure by decreasing the impact of densely local connections. Comparing the results from a series of susceptible-infected-recovered simulations on real networks, we show that our improved method can rank the nodes' spreading ability more accurately than other ranking measures such as k-shell, distance based method, mixed degree decomposition and coreness centrality method.*

*PACS: 89.75.Fb, 89.75.Hc, 89.75.Kd* DOI: 10.1088/0256-307X/33/2/028901

The study of spreading process on complex networks has attracted many researchers from congestion control to information or disease diffusion in recent decades.[1−10] One of the basic questions in understanding and controlling the spreading process is how to effectively identify influential spreaders in the network.[11−13] Many centrality indices have been proposed to address this problem, such as degree, betweenness,[14] closeness,[15] eigenvector,[16] semi-local,[17] PageRank,[18,19] and LeaderRank.[20,21] However, designing an effective method to evaluate the node spreading ability is still a large challenge up to now. In Refs. [22–25], researchers took the dynamic processes on networks into account. Konstantin *et al.*[22] provide a general framework where the dynamical influence is defined as a centrality measure to rank nodes' importance, and they found that nodes' influence is determined by the interplay between dynamics and network structure. Bauer *et al.*[23] presented a method by counting the number of potential infection walks. Li *et al.*[24] derive the node influence in terms of dynamical behaviors on the network via a Markov chain analysis. What is more, Lin *et al.*[25] proposed a dynamics-sensitive centrality which integrates network topological features and dynamical properties to quantify the spreading influences of nodes.

Recently, Kitsak *et al.*[26] found that the most influential spreaders are those located in the inner core of the network as identified by the $k$-shell decomposition analysis. However, some studies have indicated that the $k$-shell method could not distinguish nodes' spreading influence in the same shell[28−30] and could fail in the networks with core-like structure.[27] To overcome the shortcoming of the above $k$-shell method, Zeng *et al.*[28] proposed a mixed degree decomposition method to improve it by incorporating both the residual links and removed links with a tunable parameter $\lambda$. However, the value of parameter $\lambda$ is not general for each network, due to the fact that it needs to be adjusted for a given network structure. In consideration of the shortest distance between a target node to the node set in the network core, Liu *et al.*[29] presented an improved method to discriminate the nodes with the same $k$-shell value, while its computational complexity is relatively high. Bae *et al.*[30] pointed out that the spreading influence of a node depends on the number of connections to the core nodes in the network. They proposed a novel measure which is called the coreness centrality by summing all neighbors' $k$-shell values. However, when it is used to rank nodes, it only considers the neighbors' $k$-shell values while it ignores the effect of the topological connections among them on the node's spreading influence. Actually, densely local connections play a negative role in spreading,[31−34] Chen *et al.*[34] provide an effective and efficient ranking method called ClusterRank and shows that the spreading process initiated from the highly clustered nodes would be more likely to confine in a local region. Thanks to those previous researchers, by taking into account the network constraint coefficient,[35,36] which can comprehensively evaluate the effect of the number and the topological connections of the neighbors on node's spreading influence, we propose an improved coreness method which is the sum of the renewed neighbors' $k$-shell values by a correction function. The improved coreness centrality (IC) of node $i$ is defined as follows:

$$\mathrm{IC}(i) = \sum_{j \in \Gamma(i)} f(S_j) \cdot \mathrm{ks}(j), \tag{1}$$

where $\Gamma(i)$ is the neighbor node set of node $i$, $\mathrm{ks}(j)$ is the $k$-shell value of the node $j$. The term $f(S_j)$ represents the effect of node $j$'s network constraint coefficient on its spreading influence. The network constraint coefficient ($S$) of node $j$ is measured by

$$S_j = \sum_{k \in \Gamma(j)} S_{jk} = \sum_{k \in \Gamma(j)} \Big( \frac{1}{N(j)}$$

**Corresponding author. Email: ruanyirun@163.com

© 2016 Chinese Physical Society and IOP Publishing Ltd

$$+\sum_{q\in\Gamma(k)}\left(\frac{1}{N(j)}\frac{1}{N(q)}\right)\Big)^2,\ q\neq j, \tag{2}$$

where $S_{jk}$ denotes the contribution of node $k$ to the network constraint coefficient value of node $j$, and $N(j)$ and $N(q)$ represent the degree of node $j$ and $q$, respectively. We here adopt a simple decreasing decaying function $f(S_j)=10^{-S_j}$, which is used in Ref. [33]. We also investigate the effects of different forms of function with a tunable parameter such as $S_j^{\alpha}$ and $\alpha^{-S_j}$ on the ranking results, and we find that it adds slight values to improve the accuracy of the method on ranking nodes while it makes the analysis more complicated.

For better illustration, here we take node $i$ in Fig. 1 as an example to calculate its network constraint coefficient value $S_i$. Firstly, we calculate the partial constraint value of node $i$, called $S_{if}$, the common neighbor of nodes $i$ and $f$ are $e$ and $g$, then we obtain

$$\sum_{q\in(e,g)}\frac{1}{N(i)}\frac{1}{N(q)}=\frac{1}{5}\frac{1}{5}+\frac{1}{5}\frac{1}{5}=\frac{2}{25},$$

$$S_{if}=\left(\frac{1}{N(i)}+\sum_{q\in(e,g)}\frac{1}{N(i)}\frac{1}{N(q)}\right)=\left(\frac{1}{5}+\frac{2}{25}\right)^2.$$

In this way, the values of $S_{ig}$, $S_{ie}$, $S_{ij}$, $S_{ik}$ ($\Gamma(i)=\{j,k,e,f,g\}$) are obtained. Finally, we obtain node $i$'s constraint coefficient $S_i=S_{ie}+S_{if}+S_{ig}+S_{ij}+S_{ik}$. From the above calculation process we can see that a node with a small constraint coefficient indicates that the degree of the node is large and the connections among neighbors are sparse. Therefore the propagation process initiated from the node could be more likely to spread to a large portion of the network.

Recursively, we also proposed an extended measure (labeled $\mathrm{IC}_+$) by considering the improved neighborhood coreness values of nodes whose distance from the node is no more than two hops. The value of $\mathrm{IC}_+$ for node $i$ is defined as

$$\mathrm{IC}_+(i)=\sum_{j\in\Gamma(i)}\mathrm{IC}(j). \tag{3}$$

By simulating the epidemic spreading process on 12 real networks, we show that our method performs better than $k$-shell (ks), distance based method (kl), mixed degree decomposition (km) and coreness centrality ($C_{\mathrm{nc}}$ and $C_{\mathrm{nc}+}$).

To check the effectiveness of the proposed methods, 12 real networks are investigated in this work including (1) Lusseau's Bottlenose Dolphins,[37] (2) adjacency relation network of English text,[38] (3) collaboration network of network scientists,[38] (4) C. elegans metabolic networks,[39] (5) yeast protein–protein binding network,[40] (6) protein network,[41] (7) communication network of Blogs,[42] (8) western states power grid network,[43] (9) collaboration network of arxiv general relativity category (CA-GrQc),[45] (10) the router-level topology of the internet network,[27] (11) PGP network,[44] and (12) collaboration network of condensed matter physics (CA-CondMat).[45] We here select four relatively large networks (Netscience, Yeast, Protein and Blogs) as representative examples. The results of the other networks are listed in detail in Table 1.

**Table 1.** The statistical properties and ranking results of each network. Here $N$ is the number of nodes; $M$ is the number of edges; $\beta_{\mathrm{th}}$ is the epidemic threshold for a network; $\beta$ is the infection probability in SIR simulation; and $\langle\tau\rangle_{(\cdot)}$ represents the correlation of corresponding methods for given $\beta$.

| Network | $N$ | $M$ | $\beta_{\mathrm{th}}$ | $\beta$ | $\langle\tau\rangle_{\mathrm{ks}}$ | $\langle\tau\rangle_{\mathrm{km}}$ | $\langle\tau\rangle_{\mathrm{kl}}$ | $\langle\tau\rangle_{C_{\mathrm{nc}}}$ | $\langle\tau\rangle_{C_{\mathrm{nc}+}}$ | $\langle\tau\rangle_{\mathrm{IC}}$ | $\langle\tau\rangle_{\mathrm{IC}+}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Dolphins | 62 | 159 | 0.147 | 0.15 | 0.7406 | 0.8256 | 0.7407 | 0.8110 | 0.8872 | 0.8796 | 0.9283 |
| Word | 112 | 425 | 0.075 | 0.10 | 0.8369 | 0.8730 | 0.9006 | 0.8958 | 0.8874 | 0.9158 | 0.8975 |
| Netscience | 379 | 914 | 0.125 | 0.15 | 0.5508 | 0.6071 | 0.5514 | 0.6755 | 0.8223 | 0.8072 | 0.8694 |
| C.elegans | 453 | 2025 | 0.025 | 0.05 | 0.7501 | 0.7001 | 0.8153 | 0.7924 | 0.8301 | 0.8295 | 0.8531 |
| Yeast | 1458 | 1948 | 0.140 | 0.15 | 0.5575 | 0.5580 | 0.5797 | 0.7068 | 0.8229 | 0.7953 | 0.8420 |
| Protein | 2783 | 6007 | 0.065 | 0.10 | 0.5864 | 0.5280 | 0.7614 | 0.7021 | 0.8523 | 0.7745 | 0.8748 |
| Blogs | 3982 | 6803 | 0.072 | 0.10 | 0.5221 | 0.5265 | 0.5433 | 0.6634 | 0.7821 | 0.7465 | 0.8184 |
| Powergrid | 4941 | 6594 | 0.258 | 0.30 | 0.4931 | 0.5777 | 0.331 | 0.6195 | 0.7841 | 0.6970 | 0.8118 |
| CA-GrQc | 4158 | 13425 | 0.059 | 0.10 | 0.5482 | 0.5388 | 0.6632 | 0.6380 | 0.7158 | 0.6843 | 0.7302 |
| Router | 5022 | 6258 | 0.073 | 0.10 | 0.3123 | 0.3628 | 0.6118 | 0.6878 | 0.7414 | 0.7100 | 0.7766 |
| PGP | 10680 | 24316 | 0.053 | 0.10 | 0.4913 | 0.4835 | 0.6049 | 0.6663 | 0.7200 | 0.6916 | 0.7353 |
| CA-CondMat | 21, 363 | 91, 286 | 0.045 | 0.06 | 0.5700 | 0.5821 | 0.6888 | 0.6774 | 0.7653 | 0.7238 | 0.7956 |

We apply the standard susceptible-infectious-recovered (SIR)[46–48] model to evaluate the spreading influence of each node in the network. In the model, nodes belong to three possible states: (i) susceptible $S$ represents the individuals susceptible to be infected but not yet infected; (ii) infected $I$ implies individuals that have been infected and are able to pass the disease to their susceptible neighbors with probability $\beta$; and (iii) recovered $R$ denotes individuals who are infected but have recovered with probability $\lambda$, and those nodes will never be infected again. In this work, we set $\lambda=1.0$ for generality. The spreading influence initiated from node $i$ is denoted as $\sigma(i)$ by counting the number of recovered nodes over 1000 simulations for small networks and 100 for large networks.

We begin our analysis by investigating the impact of network constraint coefficient on spreading efficiency of nodes which are in the same shell. Since the spreading influence of each node is determined by the final infected population, the infection probability $\beta$ should be neither too small nor too large. If $\beta$ is too small, the spreading initiated from the target node will confine to local neighborhood nodes. On the contrary, if $\beta$ is too large, the spreading will easily reach

a large scope of the network and the role of the target node is no longer important. Thus we set the value of infection probability $\beta$ to be slightly larger than the epidemic threshold $\beta_{\rm th} \approx \langle k\rangle/\langle k^2\rangle$ of the network, where $\langle k\rangle$ and $\langle k^2\rangle$ represent the average degree and the second order average degree,[29] respectively. The value of $\beta_{\rm th}$ and $\beta$ are listed in the second and third columns of Table 1. As shown in all four real networks of Fig. 2, the nodes with the same $k$-shell value, which have a relatively high spreading influence, have a smaller value of the network constraint coefficient as expected. The reason for this result is that for different nodes in the same shell, a node with low constraint coefficient has a larger number of neighbors and the connections among them are sparse. Thus the node would have more chance to spread the information to a large scope.

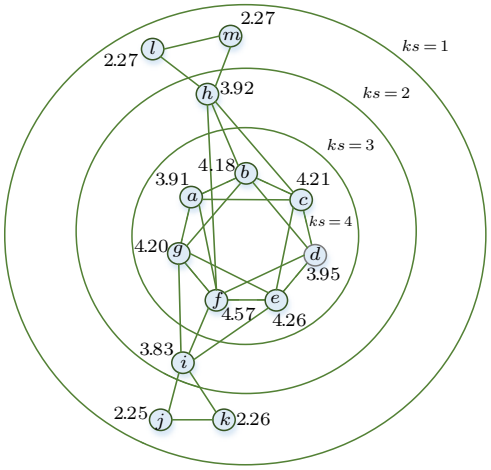


**Fig. 1.** A simple network decomposed into four layers by the $k$-shell method. There are two nodes $i$ and $h$ that locate in the same network layer, although both of them have the same coreness centrality values ($C_{\rm nc}(i) = C_{\rm nc}(h)$=16, $C_{\rm nc+}(i) = C_{\rm nc+}(h)$=70), node $i$ is expected to be of less influence than node $h$ in the epidemic spreading process since three of its neighbors, nodes $e$, $f$, $g$, link very locally within the group. Moreover, since the spreading ability of nodes $j$ and $k$ depends on node $i$ in the network, they should also be less influential than nodes $l$ and $m$, while all four nodes are considered with equal spreading ability by the neighborhood coreness method. Nodes' spreading influence is represented around each node. The results are obtained by simulating on the SIR model over 1000 independent runs with $\beta$=0.25 ($\beta_{\rm th}$=0.22). The infection result confirms our expectation.

We apply the Kendall tau coefficient $\tau$ as a rank correlation coefficient[49,50] to evaluate the performance of the ranking method, it is defined as follows:

$$\tau(R_1, R_2) = \frac{n_{\rm c} - n_{\rm d}}{\sqrt{(n_t - n_{t1})(n_t - n_{t2})}}, \qquad (4)$$

where $R_1$ and $R_2$ are two different ranking lists with $n$ elements, $n_{\rm c}$ and $n_{\rm d}$ represent the numbers of concordant and disconcordant pairs, respectively, $n_t = n(n-1)/2$, $n_{t1} = \sum_i t_i(t_i-1)/2$, $n_{t2} = \sum_j t_j(t_j-1)/2$, and $i$ and $j$ represent the number of nodes with the same rank values in the $R_1$ and $R_2$, respectively. The coefficient $\tau$ can quantify the similarity of two different ranking lists and its value is between $[-1, 1]$.

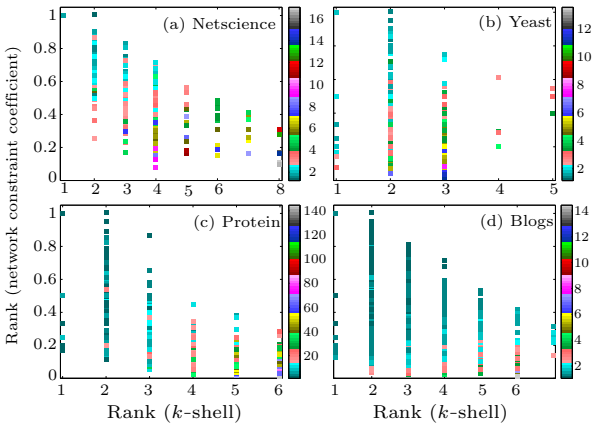


**Fig. 2.** (Color online) The networks are (a) Netscience, (b) Yeast, (c) Protein and (d) Blogs. The color of each point represents the nodes' spreading influence $\sigma$. In all four real networks, nodes which have the same $k$-shell value and relatively high spreading influence have a smaller value of the network constraint coefficient.

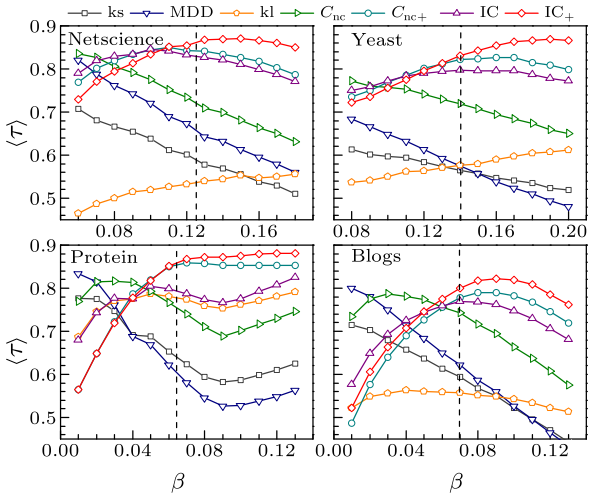


**Fig. 3.** (Color online) The value of rank correlation coefficient $\tau$ is obtained by comparing the ranking list generated by the topology-based methods and the ranking lists generated by the SIR model on networks: Netscience, Yeast, Protein and Blogs. The dotted lines correspond to the epidemic threshold $\beta_{\rm th}$ and the infection rate $\beta = (\beta_{\rm th} \pm 6\%)$ (if $\beta_{\rm th} \leq 0.07$, the spreading rate $\beta$ changes from 0.01 to 0.13).

To check the effectiveness of our method, we compare its $\tau$ values with the $\tau$ values for ks, km, kl, $C_{\rm nc}$ and $C_{\rm nc+}$ with the spreading probabilities $\beta = (\beta_{\rm th} \pm 6\%)$ (if $\beta_{\rm th} \leq 0.07$, the spreading rate $\beta$ changes from 0.01 to 0.13). As described in Fig. 3, on a wide range of probabilities $\beta$, our method presents better results than the other measures. In particular, when the infection probability $\beta$ is greater than the epidemic threshold $\beta_{\rm th}$, IC and IC$_+$ perform better than $C_{\rm nc}$ and $C_{\rm nc+}$, respectively. When $\beta$ is far smaller than $\beta_{\rm th}$, the spreading initiated from the infected node would confine to local neighbor nodes, hence the node with large degree would infect more nodes. That explains why the MDD method and $C_{\rm nc}$ index achieve larger $\langle\tau\rangle$ values when $\beta$ is small.

Sometimes, researchers are only interested in a small fraction of the most influential nodes in the network.[12] We here investigate another measurement

$\langle\tau'\rangle$ which considers a certain percentage of nodes with the strongest spreading influence measured by simulating via the SIR model. In $\langle\tau'\rangle$, we denote a percentage $L$ and consider $NL$ ($L$ from 0.05 to 1.0) nodes with the strongest spreading influence in the network, except that, the method for calculating is exactly the same as $\langle\tau\rangle$. As shown in Fig. 4, our method can achieve the best $\langle\tau'\rangle$ values on almost the entire range of $L$ in the four networks.

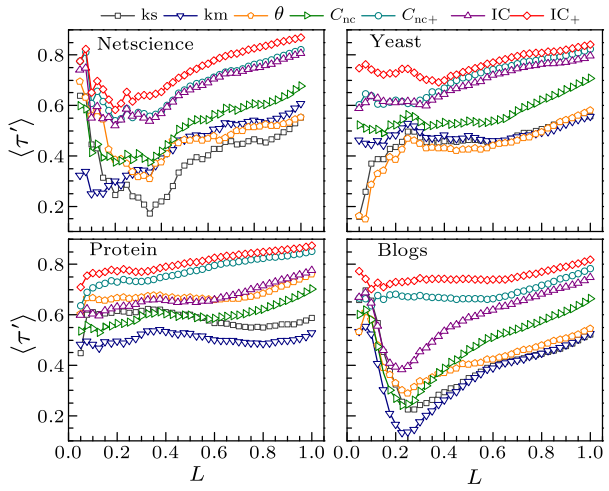


**Fig. 4.** (Color online) The values of $\langle\tau'\rangle$ for all seven methods on four real networks when $L$ varies from 0.05 to 1.0. Here $L$ represents the percentage of nodes with the largest spreading ability when calculating $\langle\tau'\rangle$. The results are obtained on the SIR model over 1000 independent runs with spreading rate $\beta$ shown in the fifth columns of Table 1.

Identifying influential spreaders in a complex network is of theoretical and practical significance. In this work, we propose an improved coreness method (IC and $IC_+$) which considers not only the $k$-shell values of the node's neighbors, but also the topological connections among them. By integrating the node's constraint coefficient which can comprehensively evaluate the number and the topological connections of the neighbors of the node, we construct a correction function to renew its neighbors' $k$-shell values. We take the final infected population in the SIR model to estimate the spreading ability of nodes in the network. The simulation results suggest that, for nodes in the same shell, the node with a smaller value of the network constraint coefficient will have a relatively high spreading ability. Further, by computing Kendall's tau coefficient $\tau$, we demonstrate that our methods can achieve better performance in most cases. Moreover, by considering only $NL$ nodes with strongest spreading ability in the network, we show that our methods work more effectively in ranking the most influential nodes than other measures.

Our results could shed some light on how to utilize the network constraint coefficient to decrease the effect of highly clustered nodes on the nodes' spreading ability. Some extensions may be made based on the network constraint coefficient. For example, by defining the combination of the degree and strength as the weighted degree of a node in weighted networks, Garas *et al.* have proposed a new weighted $k$-shell decomposition method.[51] Therefore, once the new degree value for each node in the weighted network is assigned, the network constraint coefficient can be simply generalized to weighted networks.

# References


[1] Anderson R M et al 1991 *Infectious Diseases of Humans: Dynamics and Control* (Oxford: Oxford university press)
[2] Hou L L et al 2015 *Phys. Lett.* A **379** 1321
[3] Zhou T et al 2006 *Phys. Rev.* E **74** 056109
[4] Hu Y Q et al 2014 *Phys. Rev.* X **4** 021031
[5] Pastor-Satorras R and Vespignani A 2001 *Phys. Rev. Lett.* **86** 3200
[6] Lü L Y et al 2011 *New J. Phys.* **13** 123005
[7] Zhang Y C et al 2011 *Acta Phys. Sin.* **60** 050501 (in Chinese)
[8] Zhou T et al 2006 *Prog. Nat. Sci.* **16** 452
[9] Pastor-Satorras R et al 2015 *Rev. Mod. Phys.* **87** 925
[10] Pei S and Makse H A 2013 *J. Stat. Mech.* **2013** P12002
[11] Borge-Holthoefer J and Moreno Y 2012 *Phys. Rev.* E **85** 026116
[12] Borge-Holthoefer J et al 2012 *Phys. Rev.* E **85** 066123
[13] Liu J G et al 2013 *Acta Phys. Sin.* **62** 178901 (in Chinese)
[14] Freeman L C 1978 *Soc. Networks* **1** 215
[15] Sabidussi G 1966 *Psychometrika* **31** 581
[16] Katz L 1953 *Psychometrika* **18** 39
[17] Chen D B et al 2012 *Physica* A **391** 1777
[18] Brin S and Page L 1998 *Comput. Networks ISDN Syst.* **30** 107
[19] Radicchi F et al 2009 *Phys. Rev.* E **80** 056103
[20] Lü L Y et al 2011 *PLoS One* **6** e21202
[21] Li Q et al 2014 *Physica* A **404** 47
[22] Klemm K et al 2012 *Sci. Rep.* **2** 292
[23] Bauer F and Lizier J T 2012 *Europhys. Lett.* **99** 68007
[24] Li P et al 2012 *Chin. Phys. Lett.* **29** 048903
[25] Lin J H et al 2015 arXiv:1504.06672 [cs.SI]
[26] Kitsak M et al 2010 *Nat. Phys.* **6** 888
[27] Liu Y et al 2015 *Sci. Rep.* **5** 9602
[28] Zeng A and Zhang C J 2013 *Phys. Lett.* A **377** 1031
[29] Liu J G et al 2013 *Physica* A **392** 4154
[30] Bae J and Kim S 2014 *Physica* A **395** 549
[31] Eguiluz V M and Klemm K 2002 *Phys. Rev. Lett.* **89** 108701
[32] Petermann T and Rios P 2004 *Phys. Rev.* E **69** 066116
[33] Zhou T et al 2005 *Phys. Rev.* E **71** 046141
[34] Chen D B et al 2013 *PLoS One* **8** e77455
[35] Burt R S 2009 *Structural Holes: the Social Structure of Competition* (London: Harvard University Press) p 53
[36] Burt R S et al 2013 *Annu. Rev. Psychol.* **64** 527
[37] Lusseau D et al 2003 *Behav. Ecol. Sociobiol.* **54** 396
[38] Newman M E J 2006 *Phys. Rev.* E **74** 036104
[39] Duch J and Arenas A 2005 *Phys. Rev.* E **72** 027104
[40] Jeong H et al 2001 *Nature* **411** 41
[41] Rual J F et al 2005 *Nature* **437** 1173
[42] Xie N 2006 *Master Thesis* (Bristol: University of Bristol)
[43] Watts D J and Strogatz S H 1998 *Nature* **393** 440
[44] Boguna M et al 2004 *Phys. Rev.* E **70** 056122
[45] Dorogovtsev S N et al 2006 *Phys. Rev. Lett.* **96** 040601
[46] Newman M E J 2002 *Phys. Rev.* E **66** 016128
[47] Albert R et al 2000 *Nature* **406** 378
[48] Castellano C and Pastor-Satorras R 2010 *Phys. Rev. Lett.* **105** 218701
[49] Kendall M G 1945 *Biometrika* **33** 239
[50] Knight W R 1966 *J. Am. Stat. Assoc.* **61** 436
[51] Garas A et al 2012 *New J. Phys.* **14** 083030