Physica A 391 (2012) 4012–4017

Contents lists available at SciVerse ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Identifying all-around nodes for spreading dynamics in complex networks

Bonan Hou [a,*], Yiping Yao [a], Dongsheng Liao [b]

[a] *School of Computer Science, National University of Defense Technology, 410073, Changsha, PR China*

[b] *School of Humanities and Social Sciences, National University of Defense Technology, 410073, Changsha, PR China*

ARTICLE INFO

Article history:
Received 29 September 2011
Received in revised form 23 February 2012
Available online 5 March 2012

Keywords:
Influential nodes
Ranking measures
All-around node
Complex networks

ABSTRACT

Identifying the most influential nodes in complex networks provides a strong basis for understanding spreading dynamics and ensuring more efficient spread of information. Due to the heterogeneous degree distribution, we observe that current centrality measures are correlated in their results of nodes ranking. This paper introduces the concept of all-around nodes, which act like all-around players with good performance in combined metrics. Then, an all-around distance is presented for quantifying the influence of nodes. The experimental results of susceptible-infectious-recovered (SIR) dynamics suggest that the proposed all-around distance can act as a more accurate, stable indicator of influential nodes.

1. Introduction

Identifying the most influential nodes in complex networks is an important issue for more efficient spread of information or optimal design of resource allocation [1,2]. It shows new insights for applications such as finding social leaders [3,4], influential directors [5], designing viral marketing strategies [6,7], protecting critical regions from intended attacks [8,9], and ranking reputation of publications, scientists [10,11] and athletes [12]. The topology structure plays a vital role in a network's function and behavior, thus it is reasonable to rank the nodes according to their function in the network. Due to the heterogeneous topology of most real-world networks, nodes are often endued with different roles, which are determined by both local and global topological factors. In fact, not all nodes are equal. Even those with the same degree may function differently in spreading dynamics. This makes the problem of finding influential nodes a difficult task.

The basic assumption is that given a specific spreading scheme, we rank the nodes according their impact on the range and speed of the spreading. The straightforward way is to greedily evaluate the size of the outbreak for each node or combination in the network [13,14]. This approach, however, is so computationally intensive that one may not get the result in a reasonable time. This has led to a lot of works focusing on ranking measures, which are supposed to provide objective and quantitative measures of nodes' importance, from the view of the structural analysis view. It is very common to rank the nodes according to their centrality, which measures how central a node is positioned in a network. For instance, it is widely believed that the most connected nodes (hubs) are key players in the spreading process and also the nodes with higher betweenness centrality, which measures how many shortest paths cross through this node. The widely used measures of centrality include degree, betweenness, closeness, and eigenvector centrality [15,16]. In addition, some network-based diffusion algorithms are also used to rank the nodes by taking advantage of the global features of the network. This category contains various ranking algorithms such as PageRank [17], LeaderRank [3], HITS scores [18], etc. The basic idea behind

[*] Corresponding author.
E-mail address: pipinan@gmail.com (B. Hou).

0378-4371/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2012.02.033

Table 1
The network datasets.

Network	Vertices	Links	γ	$k_{\max}$
Enron email	36,692	183,831	1.97	1383
HepPh	34,546	420,921	3.5	846
Epinions	75,879	405,740	1.69	3044
Slashdot0902	82,168	582,533	1.83	2553

these ranking algorithms is that the more frequent the node visited by the diffusion process is, the more important that node would be.

The blooming ranking measures provide a basis for better understanding the individual's function, but in some cases they may conflict with each other due to different understandings of the concept of importance or influence. For instance, more and more works find that there are plausible circumstances under which the highly connected hubs or the highest-betweenness nodes may not the most important [19,20] or influential [13,14]. Surprisingly, it has found that the connectors or bridge in the network might be a better indicator of their importance than degree centrality [21,22].

Here, we investigate the ranking measures for the node's role, and especially we are interested in the all-around nodes, those who perform well (but may not the best in sole) in different ranking metrics at the same time. First, we conduct an empirical analysis on the rank correlation among different ranking metrics with real-world networks. Then, we introduce the concept of all-around nodes in complex networks and propose a combined heuristic of all-around distance to identify them. The heuristic takes into account the nodes' degree, betweenness and core position in the network. Finally, in order to evaluate the role of all-around nodes, we apply the susceptible-infectious-recovered (SIR) spreading dynamics to quantify their impact on the network's structure and behavior. At the same time, the proposed all-around distance is compared with other commonly used ranking measures.

The remainder of this paper is organized as follows. We first analyze the rank correlation of current heuristics for node ranking (Section 2). In Section 3 we introduce the concept of all-around nodes and the all-around distance heuristic. We evaluate the role of all-around nodes and compare them with other ranking measures in Section 4. Finally, in Section 5 we expose the conclusions of the work.

2. Dataset and rank correlation

We investigate four real-world networks that represent archetypical examples of complex structures. The dataset contains (1) the Enron email network [23] covers all the communication from around half million emails, (2) the HepPh citation network within the high energy physics academic community, (3) the Epinions who-trust-whom online social networks [24] and (4) the friend/foe network of the users of the Slashdot website. These networks are characterized with heavy-tailed degree distributions that can be fitted as power-law distribution $p(k) \sim k^{-\gamma}$. For simplicity, these networks are treated as non-directed and non-weighted networks in this work. The basic information about the dataset is listed in Table 1.

We rank the nodes of these networks respectively with different ranking measures, which provide quantitative measures of the relative importance of nodes. The simplest heuristic is the degree centrality, in which the nodes with larger connectivity are considered more important. For the scale-free networks, there are a few hubs with mass links and a large portion of nodes with few links. Due to the vital role the hubs played in the network, in most cases, the degree is a powerful index for ranking the node's importance with advantage of low computational cost. Thus we take the degree ranking as the basis for comparison with other measures. They include betweenness centrality, eigenvector centrality, PageRank, HITS, and k-core. The betweenness and eigenvector centralities rank the nodes with scores of shortest-path across and eigenvector value, respectively. The PageRank and its variant HITS, widely used for web pages ranking in Wide World Web (WWW), are diffusion-based algorithms of ranking nodes by global features. And the k-core identifies the nodes' location in the core or periphery of the network according to the successive layers of the network decomposed by the k-shell decomposition method [25]. One drawback of the k-shell method is that it treats the nodes of the same layers as equals without differentiation. Because the k-core can be calculated through iteratively removing the nodes with degree $\leq k$, here we take the reverse removal order as their ranks. These ranking measures are not related with each other in essence, but in the empirical study we will see their ranking results statistically correlated in some degree. Fig. 1 takes the Zachary karate club network (in the k-shell layers layout) to illustrate the difference of these ranking measures.

Here, we analyze rank correlation for these ranking measures that are consistent with the degree centrality. We conduct this statistical analysis by using Kendall's tau method [26,27], which measures the similarity of orderings of two independent data when ranked by each of the quantities. The Kendall τ coefficient of two ranked lists, r_a and r_b, is defined as $\tau(r_a, r_b) = \frac{n_c - n_d}{n_c + n_d}$, where n_c and n_d are the number of concordant or discordant pairs, respectively. A pair of entries (o_1, o_2) are said to be concordant if o_1 appears either before or after o_2 in both lists r_a and r_b, and is called discordant if the order is reversed. The value of Kendall's correlation coefficient τ lies between -1 and $+1$, where complete correlation ($\tau = 1$) indicates that the rank results by two ranking algorithms are identical, and complete anticorrelation ($\tau = -1$) indicates that the ranks are reversed.

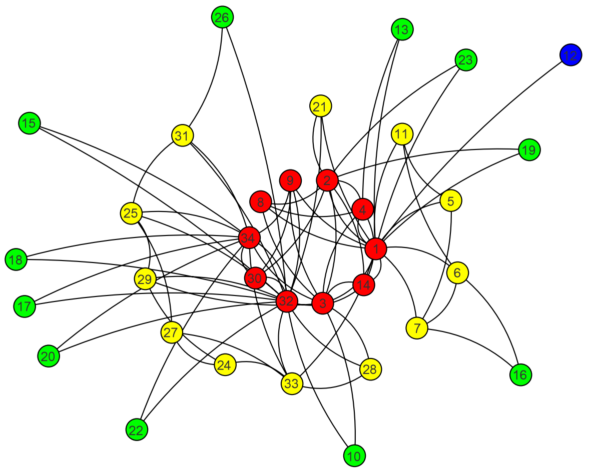
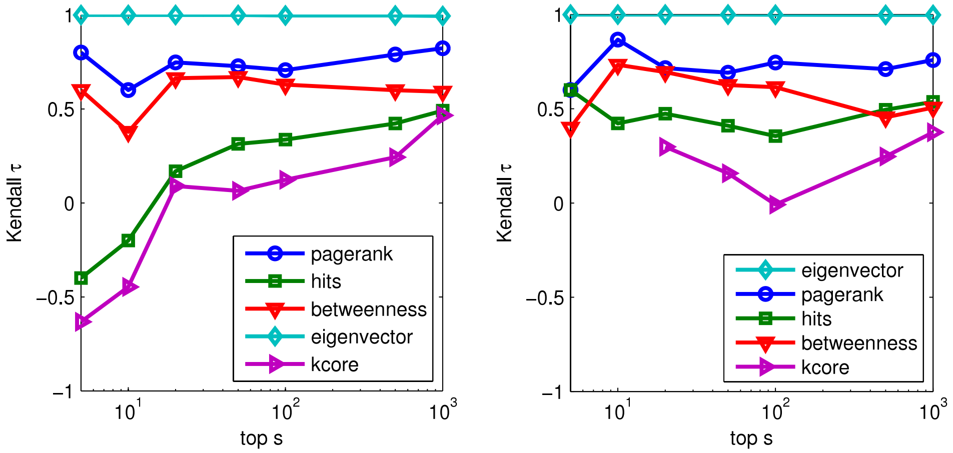


Fig. 1. The Zachary karate club network in the k-shell layers view. The k-shell method identifies nodes 1, 2, 3, 4, 8, 9, 14, 30, 32, 24 as the most inner cores with $k_s = 4$. The node 32 has highest degree, also ranked as the top by PageRank and HITS, and node 1 is of the highest betweenness centrality.



(a) Enron email network.

(b) Epinions social network.

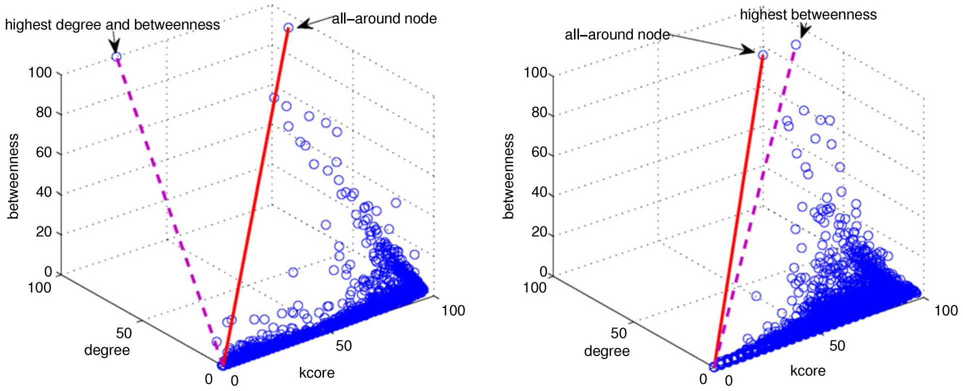
Fig. 2. The Kendall rank coefficient of different ranking measures compared with degree ranking. JUNG 2.0 library is used to calculate the degree, betweenness, eigenvector centralities, PageRank and HITS ranking. The jumping probability of PageRank we use is 0.15. In HITS, we choose the authority as nodes' scores.

We rank the nodes of the above networks with different measures. The nodes are ranked in descending order of their scores. For those with the same score, their relative order is random. Because we are often interested in the nodes listed as important ones [28], each time we truncate the top s ranks and calculate their correlation coefficient consistent with degree ranking. The empirical results on Enron email and Epinions social network are shown in Fig. 2. It indicates that the eigenvector, PageRank, and betweenness ranking measures have a positive correlation pattern with degree ranking. That means a large portion of the node ordering by these ranking measures is reserved with their degree sorting. The hubs in general can also get a relatively high score in other centralities and diffusion-based ranking algorithms. The k-shell method is an exception that is not highly correlated with degree ranking. That's partly because the hubs in the periphery of network would be possibly assigned with a low k-core k_s index.

The correlation existing in different ranking measures indicates that there is a large possibility that some nodes would gain high scores by all measures. This observation provokes us to ask whether there are nodes which would be ranked top by combined measures, and what roles they play in the network's function.

3. All-around nodes

In this section, we introduce the concept of all-around nodes and quantify them with a vector distance. We refer to the nodes whose scores are high in all different ranking measures as the all-around nodes. They like the all-around players in the arena, even if they may not be the best in an individual event, they will perform well in the combined events. So it is natural to define a combined heuristic from different aspects.



(a) Enron email network.

(b) HepPh citation network.

Fig. 3. The distribution of nodes on three dimensions of all-around measures. The node with highest degree, betweenness or *k*-core may not be the most all-around node.

Because the ranking algorithms use their own criterion to assign numerical scores to each node, the scores are relative rather than absolute, usually without a unified quantity. Thus a node's scores assigned by different metrics cannot be compared straightly. In order to evaluate the node's importance from a combined perspective, we first apply normalization techniques on the scores to the same scale but reserve their distribution on the rank list. Therefore the scores of different measures can be equally added or compared. For example, we can map the degree value onto the range 0–100 and keep the degree distribution unchanged.

Then, we define the all-around nodes with Euclidean distance as

$$d = \sqrt{\|k\|^2 + \|C_B\|^2 + \|k_s\|^2} \tag{1}$$

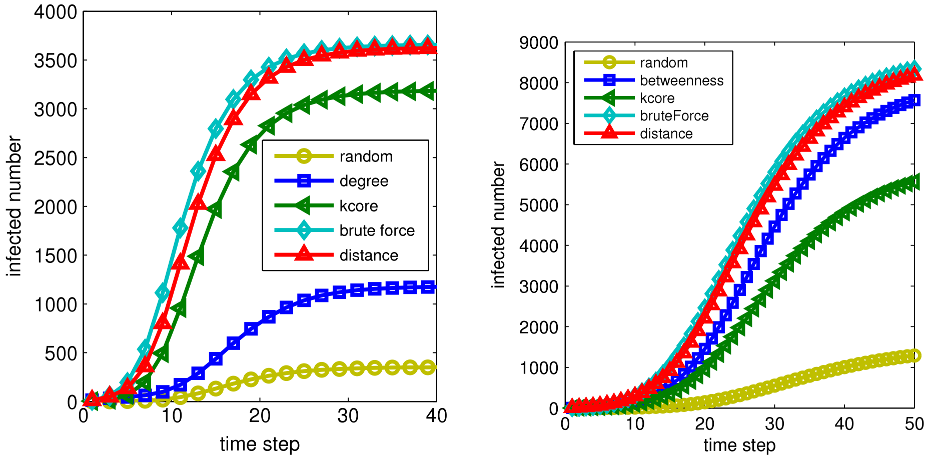
where we choose degree (k), betweenness centrality (C_B) and k-core (k_s) rankings as a composition to quantify the all-around nodes. The reasons we choose these three measures include: (1) they evaluate the node's importance from three different, representative aspects. The degree depicts the node from its local connectivity, while the betweenness counts the critical path across with global information. The k-core identifies the nodes' position from the center and periphery of the network; (2) they are less interrelated as shown by Fig. 2, so that the all-around distance can avoid in some extent the bias to the hubs. At the same time, the nodes identified as important ones by previous measures have little chance to be lagged behind in this new combined measure; (3) Except for betweenness centrality having a computational complexity of $O(nm)$ [29] (where n and m are numbers of nodes and links in the network, respectively), others can be calculated in linear time (k-shell decomposition is $O(m)$), so the total running complexity of all-around distance is $O(nm)$.

After plotting the nodes of networks onto these three dimensions, it is easy to observe the distribution of all-around nodes, illustrated by Fig. 3. In general, most nodes are aggregated with moderate distance, and a few nodes are scattered sparsely with longer distance. In the Enron email network, the node with both highest degree and betweenness is not the top all-around node, but is listed as second because the dashed line is shorter than the solid one. In the HepPh citation network, the top all-around node is with the highest degree, but not with the highest betweenness, as illustrated by Fig. 3(b). However, the most all-around node in the Epinions network coincidentally has the highest degree, betweenness and k-core. In the Slashdot network, the most all-around node is the one with highest degree, betweenness, but not inner k-core.

4. Finding influential nodes in SIR model

In this section, we study the implication of all-around nodes for finding the influential spreaders in complex networks. In the experiment, we apply the susceptible-infectious-recovered (SIR) model on the above networks. The SIR model has been widely used to describe disease, information and rumor spreading process in a population. At each time step, each infected node will infect a susceptible neighbor with probability β once interacted, and the infected will recover with rate λ. In our study we use relatively small values for β, because the large β value will make the spreading cover almost all the network, in which the role of individual is no longer important. In order to find the influential nodes, we simulate the spreading process with different setting of initial infected origins. We concern the evolution and final infected number of the population caused by the cascaded spreading process.

First, we compare the spreading by initiating only one node listed as the most important by different ranking measures. Besides the degree, betweenness, k-core and all-around distance, we also consider the randomly selected seed and the most influential node found by the brute force approach for comparison. In the random case, a node is randomly selected as a seed in each independent simulation run. In addition, we find the most influential node by running SIR model from each node in the network, and choose the node with the largest size of the outbreak. This brute force approach, however, is very time-consuming and often inapplicable in practice. Fig. 4 compares the average infected numbers of nodes who have ever infected



(a) Enron email network.

(b) HepPh citation network.

Fig. 4. Time evolution of SIR model with one node on the top of ranking by different measures initially activated. The parameters are $\beta = 0.01$ and $\lambda = 0.3$. Each point represents the average value obtained from 1000 independent simulation runs. In order to reduce the execution time of brute force approach, we first run the model from each node with 10 times and pick the top one, then run 1000 times.

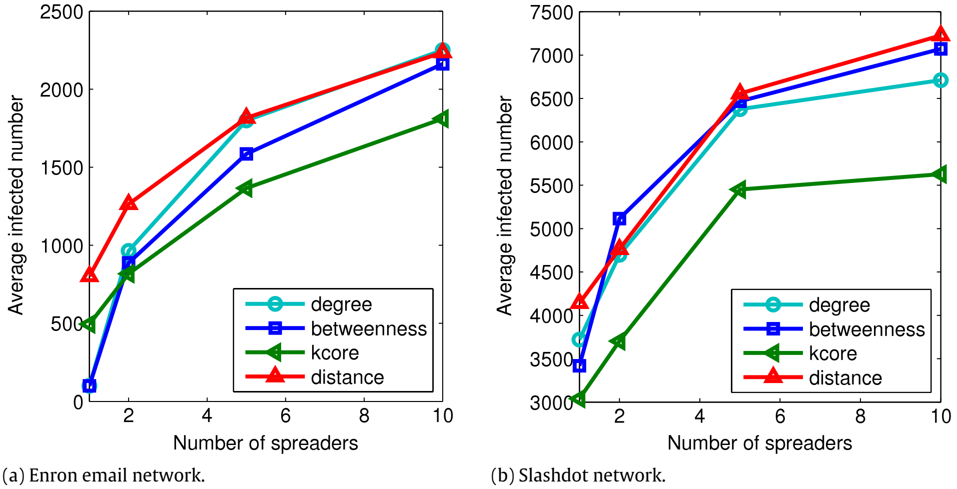
evolving along the simulation time step. On the Enron email network case (see Fig. 4(a)), the node with highest degree and betweenness fails to be the most influential one, which is contracted with our intuition. Although the top all-around node is also not the most influential one found by the costly brute force approach, their capabilities for the spreading are very close. At the same time, the top all-around node is more influential than other measures, even the k-core heuristic proposed by Kitsak et al. [1]. Although the top all-around node is positioned in the most inner core of the email network, it cannot be easily discriminated by k-shell algorithm from mixing with the other 275 nodes in the same layer. The most important node ranked by the PageRank is also the node with highest degree, but not the top all-around node. In the HepPh citation network (see Fig. 4(b)), the top all-around node is just the one with highest degree, but not the one with highest betweenness. In our experiment of the spreading process, the all-around node is still more influential than those with highest betweenness or k-core scores. And the outbreak caused by the all-around node keeps very close to the most influential one found by the brute force approach. The above results suggest that the heuristic of all-around distance can effectively find the more influential nodes.

Next, we consider the extent of spreading that starts from multiple origins simultaneously with the same model. In the simulation, the top s nodes ranked by different measures are initially activated. For simplicity, we just select the seeds from the ranking list with descending order, but ignore the connectivity information between them. In the k-core case, the seeds are randomly selected from the nodes in the most inner layer. We compare the infected numbers at step t in the simulation. In this study we use relatively small values for t, because the outbreak will become more irrelevant with the originals when t increases. Fig. 5 shows the average infected numbers on the networks when the outbreak simultaneously starts from the top s nodes with the highest k, C_B, k_s and all-around distance. The results show that the all-around distance can conduce to a larger outbreak on the email network with different number of initial spreaders. On the Slashdot network, the distance heuristic works best at $s = 1, 5$ and 10, and as a whole it is still better than degree and betweenness centralities. The results on Epinions and HepPh networks are roughly similar but not shown in the figure. In the above cases the outbreak effect becomes more subtle as s increases, part of the reason being that the cascading effect becomes more unrelated with initial conditions. It's also necessary to mention that the chosen top s nodes from different measures also share a large correlation as observed in Section 2, which means some nodes are chosen at the same time by these measures.

5. Conclusion

Due to the stochasticity of the spreading process and topological complexity of the network, it is a difficult task to characterize the nodes which are more influential than others. In this paper, we propose a ranking measure from a structural view for identifying influential nodes in complex networks. The proposed all-around distance provides a heuristic for ranking nodes through synthetically combining the degree, betweenness and k-shell ranking measures. When applied to the real-world networks, it effectively finds the more influential spreaders than other ranking measures.

Because current ranking measures may fail to find the most influential nodes in sole, we suggest that the all-around distance could be a more effective and stable indicator. Even the all-around nodes may not gain highest score in some heuristics, their comprehensive strength reserve the reason to be considered. At the same time, the nodes on the top of a single ranking have little chance to be omitted by the all-around distance heuristic, because they are more likely ranked as important by other measures.



(a) Enron email network.

(b) Slashdot network.

Fig. 5. The outbreaks with initially activating the top *s* nodes of different measures. Each point represents the average value obtained from 100 independent simulation runs.

Acknowledgments

We thank the anonymous reviewers for valuable comments on this paper. This work was funded by the National Natural Science Foundation of China (NSFC) Grant (No. 61170048, 91024030).

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] G. Ghoshal, A.-L.L. Barabási, Ranking stability and super-stable nodes in complex networks, Nature communications 2 (2011) 1–7.

[3] L. Lü, Y.-C. Zhang, C.H. Yeung, T. Zhou, Leaders in social networks, the *delicious* case, PLoS ONE 6 (6) (2011) e21202.

[4] D. Chen, L.L.M.-S. Shang, Y.-C. Zhang, T. Zhou, Identifying influential nodes in complex networks, Physica A: Statistical Mechanical and its Applications 391 (4) (2012) 1777–1787.

[5] X. Huang, I. Vodenska, F. Wang, S. Havlin, H.E. Stanley, Identifying influential directors in the united states corporate governance network, Physical Review E 84 (2011) 046101.

[6] P. Domingos, M. Richardson, Mining the network value of customers, in: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '01, ACM, New York, NY, USA, 2001, pp. 57–66.

[7] M. Richardson, P. Domingos, Mining knowledge-sharing sites for viral marketing, in: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '02, ACM, New York, NY, USA, 2002, pp. 61–70.

[8] D.R. Wuellner, S. Roy, R.M. D?Souza, Resilience and rewiring of the passenger airline networks in the United States, Physical Review E-Statistical, Nonlinear and Soft Matter Physics 82 (5 Pt 2) (2009) 11.

[9] R. Albert, H. Jeong, A.-L. Barabasi, Error and attack tolerance of complex networks, Nature 406 (6794) (2000) 14.

[10] F. Radicchi, S. Fortunato, B. Markines, A. Vespignani, Diffusion of scientific credits and the ranking of scientists, Physical Review E 80 (2009) 056103.

[11] P. Chen, H. Xie, S. Maslov, S. Redner, Finding scientific gems with Google's pagerank algorithm, Journal of Informetrics 1 (1) (2007) 8–15.

[12] F. Radicchi, Who is the best player ever? a complex network analysis of the history of professional tennis, PLoS ONE 6 (2) (2011) e17249.

[13] D. Kempe, J. Kleinberg, E. Tardos, Maximizing the spread of influence through a social network, in: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '03, ACM, New York, NY, USA, 2003, pp. 137–146.

[14] W. Chen, Y. Wang, S. Yang, Efficient influence maximization in social networks, in: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '09, ACM, New York, NY, USA, 2009, pp. 199–208.

[15] D. Kosch, K.A. Lehmann, L. Peeters, S. Richter, Centrality indices, Network 3418/2005 (2005) 16–61.

[16] K. Okamoto, W. Chen, X.-Y. Li, Ranking of closeness centrality for large-scale social networks, in: Proceedings of the 2nd Annual International Workshop on Frontiers in Algorithmics, FAW '08, Springer-Verlag, Berlin, Heidelberg, 2008, pp. 186–195.

[17] S. Brin, L. Page, The anatomy of a large-scale hypertextual web search engine, Computer Networks and ISDN Systems 30 (1998) 107–117.

[18] J.M. Kleinberg, Authoritative sources in a hyperlinked environment, in: Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '98, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1998, pp. 668–677.

[19] H. Jeong, S.P. Mason, A.-L. Barabasi, Z.N. Oltvai, Lethality and centrality in protein networks, Nature 411 (6833) (2001) 41–42.

[20] Y.-Y. Liu, J.-J. Slotine, A.-L. Barabasi, Controllability of complex networks, Nature 473 (7346) (2011) 167–173.

[21] R. Guimera, L.A.N. Amaral, Functional cartography of complex metabolic networks, Nature 433 (7028) (2005) 895–900.

[22] J.-D.J. Han, N. Bertin, T. Hao, D.S. Goldberg, G.F. Berriz, L.V. Zhang, D. Dupuy, A.J.M. Walhout, M.E. Cusick, F.P. Roth, et al., Evidence for dynamically organized modularity in the yeast protein–protein interaction network, Nature 430 (6995) (2004) 88–93.

[23] J. Leskovec, K.J. Lang, A. Dasgupta, M.W. Mahoney, Community structure in large networks: natural cluster sizes and the absence of large well-defined clusters, Internet Mathematics 6 (1) (2008) 66.

[24] M. Tahajod, A. Iranmehr, N. Khozooyi, Trust management for semantic web, Computer and Electrical Engineering, International Conference on 2, 2009, pp. 3–6.

[25] S. Carmi, S. Havlin, S. Kirkpatrick, Y. Shavitt, E. Shir, A model of internet topology using *k*-shell decomposition, Proceedings of the National Academy of Sciences 104 (27) (2007) 11150–11154.

[26] E.L. Lehmann, Nonparametrics: Statistical Methods Based on Ranks, Holden-Day, 1974.

[27] M.G. Kendall, A new measure of rank correlation, Biometrika 30 (1/2) (1938) 81–93.

[28] S. Fortunato, M. Boguna, A. Flammini, F. Menczer, How to make the top ten: approximating pagerank from in-degree, Nov. 2005, arXiv:cs.IR/0511016.

[29] U. Brandes, A faster algorithm for betweenness centrality, Journal of Mathematical Sociology 25 (2) (2001) 163–177.