Identifying influential spreaders in complex networks based on entropy method

Xiaojian Ma [a], Yinghong Ma

School of Business, Shandong Normal University, Jinan 250014, China;

[a]xiaojianma0813@163.com

Abstract

It is a fundamental task in network to identifying the importance of nodes. In this paper, we propose a measure by a semi-local manner which considering both the nearest and next neighbor of a node .Under the help of entropy value, the node whoes neighbors with uniform distribution is assigned to be a high centrality value. The experimental results have showed that the proposed method can better distinguish the influence on nodes when compared with DC, BC, CC, and H-index. And the proposed method is highly correlated with the result of SIR ranking and outperforms other methods in evaluating the nodes' spreading.

Keywords

Complex network; Influential nodes; SIR epidemic model.

1. Introduction

Identifying influential nodes in complex network is of theoretical and practical significance, such as the controlling and dissemination of information. A significant number of measures have been proposed in recently years to evaluate the importance of nodes. Degree centrality (DC) [1] is a basics measures, and the importance of one node is measured by the number of neighbors. Global measures such as betweenness centrality (BC) [2] and closeness centrality (CC)[3] can identify node influences in the global scope. K-shell [4] method has been proposed to identify the importance of nodes. Later, researchers proposed other modified method based on the K-shell to further improve the ranking performance [5]. Chen proposed a semi-local centrality [6], which take more neighbors information into consideration when evaluate the nodes' importance. H-index [7] considers the importance of nodes according to the concept of h-index, that is to say there exist at least h neighbors whose degree is no less than h.

The nodes with high degree value seem to be more influential in the spreading process, since more nodes will likely to be influenced in the process. However, more adjacent information will be neglected since the degree only considers nodes in a local scope. So, an effective measure capturing more information should be employed to evaluate the nodes' importance in general information transfer. Motivated by these considerations, in this paper, we design a measure, which considers a node's centrality by taking the nearest and the next nearest neighbors into consideration, the application of entropy which achieves an effect that the nodes' whose neighbors with uniform distribution is assigned to be a high centrality value. The SIR model is employed to examine the performance of the proposed method. Experiment results of several real networks suggest the measure outperforms the other measures in terms of correctness of the ranking list. In addition, the method can be easily applied in large-scale networks since it only requires known local information.

The part of the paper is organized as followes. The details of the proposed measure are described in Section 2. The simulation strategies and experimental results are present in Section 3. Section 4 the discussions is given.

2. Proposed method

2.1 Entropy

Information entropy, is widely used in information science and statistical physics to describe the order of information distribution. The Shannon's entropy of p is defined as:

$$E(p) = -\sum_{i=1}^{n} p_i \log(p_i) \tag{1}$$

Where $p_i (i = 1, 2, ..., n)$ be a probability vector of event, and $0 \le p_i \le 1$, $\sum_{i=1}^{n} p_i = 1$.If the probabilities p_i have uniform distribution, higher the entropy value will be [8].and as the value of n increases, so does the entropy value. Therefore, the entropy method can be employed to detection nodes with more uniform neighbors[9].

2.2 Proposed method

A Social Network is denoted as an graph consisting of nodes and edges shown as $G = (V, E)$,where V represent the set of nodes in network and $E \subseteq V \times V$ represent the set of edges between nodes. The Entropy Centrality of node v_i is denoted by $C(v)$:

$$C(v) = -\sum_{w \in \Gamma_2(v)} \frac{d_w}{\sum_{w \in \Gamma_2(v)} d_w} * \log\left(\frac{d_w}{\sum_{w \in \Gamma_2(v)} d_w}\right) \tag{2}$$

$$EC(v) = \sum_{u \in \Gamma_1(v)} C(u) \tag{3}$$

Where dw is the degree of node w , and $w \in \Gamma_2(v)$ is the nearest and next nearest neighbors of node v and $u \in \Gamma_1(v)$ is the nearest neighbors of node v .

3. Experiment

3.1 Dataset

Some details about the real-world network can be found in Table 1.Including Karate club network (Karate) [10] ,Pol-book network (Polbook) [11],US air line (Usair)[12], E-mail network(Email)[13]. The details include the Node number, Edge number,average Degree, Maximum degree and Clustering coefficient.

Table 1 Some properties of network

Network	Node number	Edge number	Max degree	Average degree	Cluster
Karata	34	78	17	4.5882	0.5706
Polbook	105	441	25	8.4	0.4875
Usair	332	2126	139	12.8072	0.6252
Email	1133	5451	71	9.623	0.2219

3.2 Evaluation

In this section, the Proposed method is compared with other four well-known measures in aspect of discriminability and correctness. These measures include degree centralty[1], betweenness centrality[2], closeness centrality[3], and H-index[7].

3.2.1 Distinguishability

To evaluate the capability of different measures in distinguishing each nodes' spreading ability. In this section, Complementary Cumulative Distribution Function (CCDF) [14]is utilized to measure how well the specification of the ranking distributions by different measures, calculated by:

$$CCDF(r) = \frac{|V| - \sum_{i=1}^{r} n_i}{|V|}$$

Where $|V|$ is the total numbers of nodes and n_i denotes the number of nodes in rank i, and. A slower slope in CCDF, a better ranking distribution performance the method achieves.

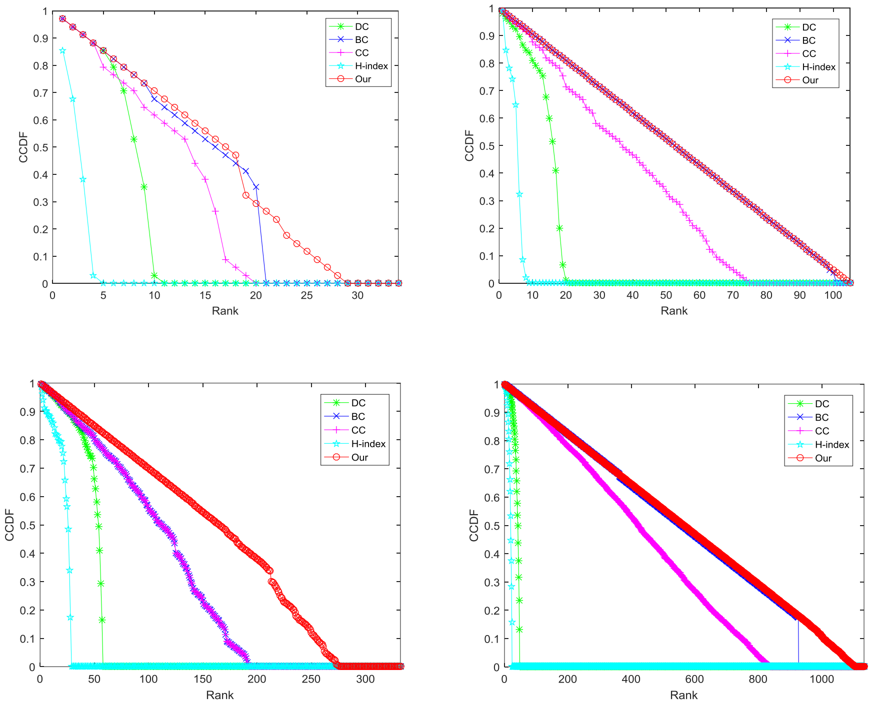


Fig.1 CCDF plots for ranking list offered by different measures

The CCDF is plotted for Karata, Polbook, Usair and Email shown in Fig. 1, AS can be seen in it, the proposed method achieve a better performance in terms of the ranking distributions. The degree method considers the neighbor information, while many nodes may have the same number of neighbors, so the degree cannot have a better performance in distinguishing the nodes' influence. The BC and CC method measure the nodes' influence in a global scope, but our method still perform better even under the circumstance that the two method consider more information.

3.2.2 Rank correlation

In the next section, the susceptible-infectious-recovered (SIR) model [15] is utilized to examine the performance of the measures. The ranking of different measures are compared with the ranking obtained from the SIR model. In the SIR model, At the beginning , one node is set to be infected, all the other nodes are in state susceptible. Then, each infected nodes infect its susceptible neighbors with a probability of β and then the node itself moves to the R state. The spreading process is repeated until no nodes with state I. And the number of the recovered nodes denotes the spreading influence of the initially infected node. The value of β should be set as larger than the threshold value β_{th} [16] , calculated as $\frac{\langle k \rangle}{\langle k^2 \rangle}$,where $\langle k \rangle$ and $\langle k^2 \rangle$ denote the average degrees and average second-order degree of the nodes ,respectively.

The kendall's τ is employed as a rank correlation coefficient [17]. By comparing the ranking list obtained from the SIR simulation with the ranking list generated through each of the measures .

$$\tau(X,Y) = \frac{N_c - N_d}{\frac{1}{2}N(N-1)}$$

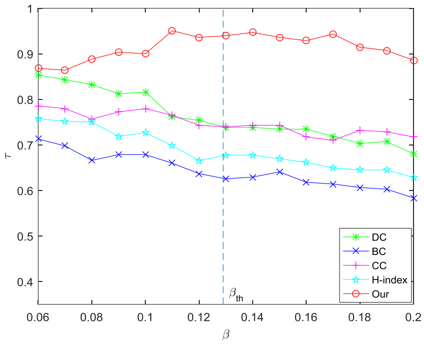
Where N_c and N_d are the numbers of concordant and discordant pairs in the ranking lists respectively the higher τ is ,the more concordant of the ranking lists with the SIR spreading .

Firstly, the rank correlations of DC, BC, CC, H-index and our proposed method are compared in Karate, Polbook, Usair and Email networks. Shown as in Table 2 the rank correlation τ under the infection probability β are calculated. The result shows that our method is more correlated with the

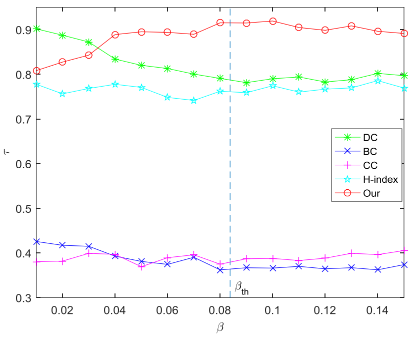
SIR spreading process. Next, the performance of different method is evaluated under different infection probability β .As can observed from Fig. 2 that the proposed method outperforms other method when the infection probability β is around the threshold value β_{th} .

Table 2 The Rank correlation

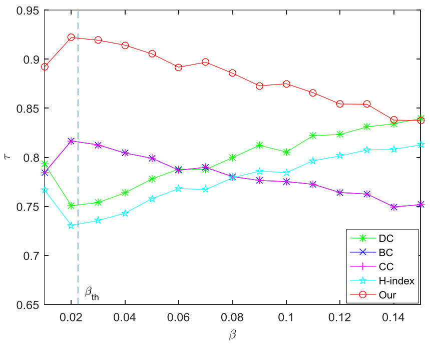
Network	β_{th}	β	$\tau(DC,\theta)$	$\tau(BC,\theta)$	$\tau(CC,\theta)$	$\tau(Hin,\theta)$	$\tau(Our,\theta)$
Karata	0.129	0.13	0.7385	0.6256	0.7395	0.6779	0.9397
Polbook	0.0838	0.09	0.7814	0.3669	0.3871	0.7588	0.9149
Usair	0.0225	0.03	0.7538	0.8126	0.8126	0.7356	0.9193
Email	0.0535	0.06	0.7913	0.6308	0.8145	0.7711	0.8351



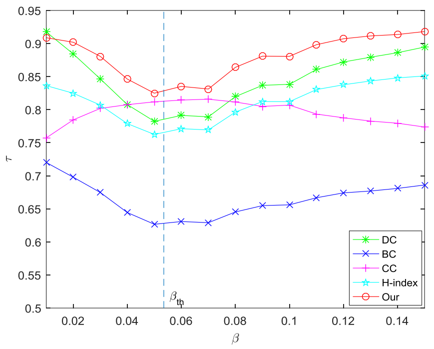
Karata



Polbook



Usair



Email

Fig. 2 Rank correlation τ under different probability β

4. Conclusion

It is a fundamental task in network to identifying nodes' importance, especially the application in epidemic spread control. In this paper, an effective measure is proposed in evaluating the influence of nodes. The nodes' importance is taken into consideration by a manner of semi-local in which both the nearest and the next nearest neighbors are taken into consideration. Through application of entropy value, the node who has neighbors with uniform distribution is assigned to be a high centrality value. The experimental results have demonstrated that the proposed method can better distinguish the influence on nodes with more nodes are assigned with different centrality values when compared with DC, BC, CC, and H-index. Another experiment shows that the proposed method is highly correlated with the result of SIR ranking.

References

[1] Freeman L C. Centrality in social networks conceptual clarification[J]. Social Networks, 1978, 1(3):215-239.

[2] Freeman L C. A Set of Measures of Centrality Based on Betweenness[J]. Sociometry, 1977, 40(1):35-41.

[3] Sabidussi G. The centrality index of a graph[J]. Psychometrika, 1966, 31(4):581-603.
[4] Kitsak M, Gallos L K, Havlin S, et al. Identification of influential spreaders in complex networks[J]. Nature Physics, 2011, 6(11):888-893.
[5] Wang Z, Zhao Y, Xi J, et al. Fast ranking influential nodes in complex networks using a k-shell iteration factor[J]. Physica A Statistical Mechanics & Its Applications, 2016, 461:171-181.
[6] D. Chen, L. Y. Lü, M. S. Shang, et al, Identifying influential nodes in complex networks, Physica A, 391(4) (2011) 1777–1787.
[7] Lü L, Zhou T, Zhang Q M, et al. The H-index of a network node and its relation to degree and coreness[J]. Nature Communications, 2016, 7:10168.
[8] COVER TM , THOMAS JA . Elements of information theory. 2nd ed; 2006 .
[9] Zareie A, Sheikhahmadi A, Fatemi A. Influential nodes ranking in complex networks: An entropy-based approach[J]. Chaos Solitons & Fractals, 2017, 104:485-494.
[10] Zachary W W. An Information Flow Model for Conflict and Fission in Small Groups[J]. Journal of Anthropological Research, 1977, 33(4):452-473.
[11] Information on http://www.orgnet.com/
[12] V1adimir Batagelj and Andrej Mrvar Pajek datasets.http://Vlado.fmf.uni-lj.si/pub/networks / data / .
[13] Guimerã R, Danon L, DãAz-Guilera A, et al. Self-similar community structure in a network of human interactions[J]. Phys Rev E, 2003, 68(6 Pt 2):065103.
[14] Bae, S. Kim, Identifying and ranking influential spreaders in complex networks by neighborhood coreness, Physica A 395 (2014) 549–559.
[15] Kitsak M, Gallos L K, Havlin S, et al. Identification of influential spreaders in complex networks[J]. Nature Physics, 2011, 6(11):888-893.
[16] Bae, J., & Kim, S. (2014). Identifying and ranking influential spreaders in complex networks by neighborhood coreness. Physica A, 395, 549–559.
[17] Nelsen RB. Kendall tau Metric, encyclopedia of mathematics. Springer; 2001. ISBN: 978-1-55608-010-4.