ARTICLE IN PRESS

Physica A ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Physica A

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

Identification of influential nodes in complex networks: Method from spreading probability viewpoint

Zhong-Kui Bao [a], Chuang Ma [a], Bing-Bing Xiang [a], Hai-Feng Zhang [a,b,*]

[a] School of Mathematical Science, Anhui University, Hefei 230601, PR China

[b] Department of Communication Engineering, North University of China, Taiyuan, Shan'xi 030051, PR China

HIGHLIGHTS

- A semi-local method is proposed to identify influential nodes in complex networks.
- The index considers main factors of spreading probability simultaneously.
- The transmission rate is approximately replaced by other features.
- The performance of our index is genernally better than other indices.

ARTICLE INFO

Article history:
Received 24 June 2016
Received in revised form 10 September 2016
Available online xxxx

Keywords:
Complex networks
Influential spreader
Reciprocal of average degree
Epidemic threshold

ABSTRACT

The problem of identifying influential nodes in complex networks has attracted much attention owing to its wide applications, including how to maximize the information diffusion, boost product promotion in a viral marketing campaign, prevent a large scale epidemic and so on. From spreading viewpoint, the probability of one node propagating its information to one other node is closely related to the shortest distance between them, the number of shortest paths and the transmission rate. However, it is difficult to obtain the values of transmission rates for different cases, to overcome such a difficulty, we use the reciprocal of average degree to approximate the transmission rate. Then a semi-local centrality index is proposed to incorporate the shortest distance, the number of shortest paths and the reciprocal of average degree simultaneously. By implementing simulations in real networks as well as synthetic networks, we verify that our proposed centrality can outperform well-known centralities, such as degree centrality, betweenness centrality, closeness centrality, k-shell centrality, and nonbacktracking centrality. In particular, our findings indicate that the performance of our method is the most significant when the transmission rate nears to the epidemic threshold, which is the most meaningful region for the identification of influential nodes.

1. Introduction

Many social, technological and biological systems can be described in terms of networks to characterize the intricate interaction topologies. The nodes in the networks often play different roles in maintaining the global functionality of the system, promoting the new product promotion and the diffusion of information, hindering the spread of epidemic and so on. Thus, how to effectively identify influential spreaders is a central topic of interest in networks [1–8]. In doing so, many

* Corresponding author at: School of Mathematical Science, Anhui University, Hefei 230601, PR China.
E-mail address: haifengzhang1978@gmail.com (H.-F. Zhang).

http://dx.doi.org/10.1016/j.physa.2016.10.086

centrality indices such as degree centrality (DC) [9], betweenness centrality (BC) [10], closeness centrality (CC) [11], k-shell method (KS) [1] and a uniform H-index [12] have been adopted to identify the most influential spreaders in networks. Degree centrality is a straightforward index, however, a node with larger degree value located in the periphery of networks may has small influence. Betweenness centrality is a global measure and is not suitable for large-scale networks due to its higher computation complexity. Closeness centrality only considers the shortest distance between pairs of nodes but ignore the number of the shortest paths. K-shell method indicates that the nodes with the highest k-core values have the most influence. However, recent researches have demonstrated that the nodes within the same shell often have distinct influences, and this method is invalid in networks without core-like structures and has low resolution. Thus, some methods were proposed to further improve the performance of the k-shell method [13–17].

From spreading viewpoint, on one hand, whether the information can easily diffuse from a spreader is not only sensitively dependent on the shortest distance from it to other nodes, but also the number of shortest paths. Take the schematic network in Fig. 1 as an example, even though the shortest distances from nodes j and k to i are the same ($d_{ij} = d_{ik} = 2$). In general, the probability of node j being infected by node i is larger than that of node k since there are three 2-hops shortest paths between i and j; On the other hand, the spreading dynamics is also closely related to the transmission rate of information (denoted by β), which is often ignored in previous literatures [18–20]. For instance, for nodes s and j in Fig. 1, the probabilities of node j and node s being infected by node i approximately equal to $1-(1-\beta^2)^3$ and β when ignoring the impacts of the external factors, respectively. Therefore the probability of node j being infected is larger than that of node s when $\beta \succeq 0.4$. Inspired by the above mentions, an ideal case is that the new proposed index should incorporate the three factors simultaneously. Recently, Liu et al. have made some attempts to consider the effect of the transmission rate on the identification of influential nodes from other perspective [21].

However, it is very hard for us to know the value of transmission rate β in advance and which is different from different categories. As we know, whether the spreading process can outbreak on networks depends on its epidemic threshold (β_c). Spreading dynamics will extinct eventually when $\beta \ll \beta_c$; On the contrary, the spreading process can propagate large fraction of networks when $\beta \gg \beta_c$. Therefore, only the region near the epidemic threshold is meaningful for the identification of influential nodes, too small value of β cannot lead to the wide diffusion of information and too large value of β causes the large-scale diffusion of information, which is independent on which one spreader is chosen [22]. It is well-known that the epidemic threshold of SIR (Susceptible–Infected–Recovery) model is $\beta_c = \frac{\langle k\rangle}{\langle k^2\rangle-\langle k\rangle}$ ($\langle k^2\rangle$ is the second moment degree of the network). In this paper, we use $\frac{1}{\langle k\rangle}$ ($\langle k\rangle$ is the average degree of the network) to approximately replace the transmission rate β in our proposed centrality index. The reason is: firstly, $\frac{1}{\langle k\rangle} > \frac{\langle k\rangle}{\langle k^2\rangle-\langle k\rangle}$ in many cases, which can ensure the spreading process can outbreak in networks; secondly, $\frac{1}{\langle k\rangle}$ is not so large that the epidemic can propagate most of nodes in general; third, by using $\frac{1}{\langle k\rangle}$ to approximate the transmission rate can avoid to know the additional information of networks, i.e., $\langle k^2\rangle$. Meanwhile, to reduce the computation complexity, we just consider the shortest distance within 3-hops, therefore, the index can be viewed as a semi-local index [23].

The layout of the paper is as follows: In Section 2, we first briefly introduce several typical centrality indices which are compared in this work, and the description of our method is also presented. Then the main results are presented in Section 3. Finally, conclusions are summarized in Section 4.

2. Method

An undirected network is represented by $G = (N, E)$ with N nodes and E edges, and its structure can be described by an adjacent matrix $A = (a_{ij})_{N\times N}$ where $a_{ij} = 1$ if node i is connected to node j, and $a_{ij} = 0$ otherwise.

Here we briefly review the definitions of several centrality indices that will be discussed in this work.

The degree centrality (DC) of node i is defined as the number of nearest neighbors [9], namely

$$DC(i) = \sum_{j=1}^{N} a_{ij}. \tag{1}$$

The betweenness centrality (BC) of node i is defined as the fraction of all shortest paths travel through the node [10], which is denoted as

$$BC(i) = \sum_{s\neq i\neq l} n_{sl}^{i}/n_{sl} \tag{2}$$

with n_{sl} and n_{sl}^{i} be the number of shortest paths between nodes s and l, and the number of shortest paths between s and l passing through node i.

Closeness centrality (CC) of node i is defined as the reciprocal of the sum of shortest distances to all other nodes [11]:

$$CC(i) = \frac{1}{\sum_{j\neq i} d_{ij}}. \tag{3}$$

Please cite this article in press as: Z.-K. Bao, et al., Identification of influential nodes in complex networks: Method from spreading probability viewpoint, Physica A (2016), http://dx.doi.org/10.1016/j.physa.2016.10.086