Physica A 452 (2016) 289–298

Contents lists available at ScienceDirect

Physica A

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

Identify influential spreaders in complex networks, the role of neighborhood

Ying Liu [a,b,c], Ming Tang [a,d,*], Tao Zhou [a,d], Younghae Do [e]

[a] Web Sciences Center, University of Electronic Science and Technology of China, Chengdu 611731, China
[b] School of Computer Science, Southwest Petroleum University, Chengdu 610500, China
[c] Section for Science of Complex System, Medical University of Vienna, Vienna 1090, Austria
[d] Big Data Research Center, University of Electronic Science and Technology of China, Chengdu 611731, China
[e] Department of Mathematics, Kyungpook National University, Daegu 702-701, South Korea

HIGHLIGHTS

- A centrality measure encoding the centrality of a node's neighborhood is proposed.
- The neighborhood centrality outperforms degree and coreness in ranking spreaders.
- A saturation effect of the neighborhood is discovered.
- Considering the 2-step neighborhood of a node balances the cost and performance.

ARTICLE INFO

Article history:
Received 2 November 2015
Received in revised form 13 January 2016
Available online 18 February 2016

Keywords:
Epidemic spreading
Influential spreader
Neighborhood centrality
Saturation effect

ABSTRACT

Identifying the most influential spreaders is an important issue in controlling the spreading processes in complex networks. Centrality measures are used to rank node influence in a spreading dynamics. Here we propose a node influence measure based on the centrality of a node and its neighbors' centrality, which we call the neighborhood centrality. By simulating the spreading processes in six real-world networks, we find that the neighborhood centrality greatly outperforms the basic centrality of a node such as the degree and coreness in ranking node influence and identifying the most influential spreaders. Interestingly, we discover a saturation effect in considering the neighborhood of a node, which is not the case of the larger the better. Specifically speaking, considering the 2-step neighborhood of nodes is a good choice that balances the cost and performance. If further step of neighborhood is taken into consideration, there is no obvious improvement and even decrease in the ranking performance. The saturation effect may be informative for studies that make use of the local structure of a node to determine its importance in the network.

1. Introduction

Identifying the most influential spreaders is an important step in promoting the adoption of new ideas, products and innovations, and preventing the epidemic disease from being pervasive. Centrality measures are used to rank the importance of a node in a network, such as degree [1], closeness centrality [2], betweenness centrality [3], PageRank centrality [4], LeaderRank centrality [5,6], eigenvector centrality [7], dynamic-sensitive centrality [8,9] and coreness centrality [10]. It

* Corresponding author at: Web Sciences Center, University of Electronic Science and Technology of China, Chengdu 611731, China.
E-mail address: tangminghan007@gmail.com (M. Tang).

http://dx.doi.org/10.1016/j.physa.2016.02.028
0378-4371/

contains the idea that there is a relationship between the topological position of a node in the network and its influence and capacity in a spreading dynamics. Among them, degree is the simplest way and is most widely used, but it is based only on the local connections of a node. Subsequent research pointed out that the coreness of node k_S as identified by the k-shell decomposition [11] is a more accurate way in ranking node influence [10], which has a low computational complexity of $O(N+E)$ [12], where N is the network size and E is the number of edges. In addition, the k_S index has a good robustness, which means that the relative ranking of the k_S value for the same node remains unchanged when the network structure is incomplete, missing even up to 50% of the edges [10].

It is pointed out that the importance of a node is not determined solely by its direct connections, but also depends on the connections of its neighbors [13]. Research results show that ranking measures by considering the neighborhood of a node are more accurate in identifying the spreading influence of nodes [14–18]. For example, by considering the number of neighbors within 4-step from the node, a local centrality measure is proposed which outperforms the centrality of degree and betweenness [14]. In a recent work, a gravity centrality is proposed which uses the gravity formula to calculate the contribution of neighbors to the centrality of the considered node. This method can effectively identify influential spreaders in real-world networks and artificial networks [18]. The sum of the coreness of the nearest neighbors of a node is a better indicator of node spreading influence than the coreness [19]. In a ranking algorithm of iterative resource allocating, by considering the centrality of neighbors in a resource allocating process, the final resource a node obtains is used to rank its spreading capability. This method shows a great improvement over degree, closeness, and betweenness [20]. In addition, there are works based on counting the number of possible infection paths [21,22] in the neighborhood to determine node influence.

Intuitively, the larger neighborhood is taken into consideration, the more accurately we can predict the spreading outcome of a node. However, in the research of spreading phenomena in social networks, such as the spread of smoking, alcohol consumption, happiness, health screening, it is discovered that on average there is a significant relationship in behaviors between a node and its direct neighbors (1-step neighbor), and up to the neighbors' neighbors' neighbors (3-step neighbor), which is called the three degree of separation [23]. This implies an influence range from the spreading origin to the affected population. In addition, it is a challenging task to collect the complete network information in some real-world networks [15], due to the large amount and the temporal and spatial change of the data, such as Twitter and Facebook. Analyzing a large neighborhood seems unfeasible in such networks.

Given the effect of neighborhood, in this paper we first propose a neighborhood centrality based on the centrality of a node and its neighbors within multiple steps. Specifically, we take the degree and coreness as the benchmark centrality to study the performance of the neighborhood centrality when 1–4 steps of neighbors are considered. We find that in general the neighborhood centrality outperforms the centrality of the node. Furthermore, in most of the studied networks, considering the neighborhood within 2-step will result in a good neighborhood centrality. When the considered steps is greater than 2, the improvement of ranking accuracy is not obvious and even negative. This means a saturation effect of the neighborhood on a node. Discovering the saturation effect is meaningful in that when we consider the effect of neighborhood, taking the 2-step neighborhood into account will balance the cost and effect.

2. Methods

In this part, we first introduce the centrality measures of degree and coreness, which are used as the benchmark centrality. Then we propose the neighborhood centrality based on the degree or coreness of a node and its neighbors. Finally we describe the susceptible–infected–recovered (SIR) model used in the spreading process, and give a brief description of the data sets used in the study.

2.1. Degree centrality

The degree centrality of a node i is defined as $k_i = \sum_{j\in G\setminus i} a_{ij}$, where j is a node in the network G, and $a_{ij} = 1$ if there is a link between node i and node j, otherwise $a_{ij} = 0$. Degree is the simplest centrality measure in quantifying node importance. The larger the degree, the more neighbors the node is able to influence directly.

2.2. Coreness centrality

The coreness centrality of a node is obtained in the k-shell decomposition process. The k-shell decomposition method is used to decompose the network into hierarchically ordered shells from the core to the periphery. Initially, nodes with degree $k = 1$ are removed from the network together with their links. After removing all nodes with $k = 1$, there may appear some nodes with only one link left. We continue to remove these nodes until there is no node with one link left in the network. The removed nodes are assigned with an index $k_s = 1$. Next, nodes with degree $k \leq 2$ are removed in a similar way and assigned with an index $k_s = 2$. The pruning process continues removing higher shells until all nodes are removed. As a result, each node is assigned with a k_s value, which is called the coreness of a node. The coreness reflects the location importance of a node in the network. A large k_s means a core position in the network, while a small k_s defines the periphery of the network. Coreness centrality is considered to be a better measure than degree to identify the influential spreaders in a network [10,24].