Physics Letters A 381 (2017) 976–983

Contents lists available at ScienceDirect

Physics Letters A

www.elsevier.com/locate/pla

PHYSICS LETTERS A

Identifying multiple influential spreaders by a heuristic clustering algorithm

Zhong-Kui Bao [a], Jian-Guo Liu [b], Hai-Feng Zhang [a,c,*]

[a] School of Mathematical Science, Anhui University, Hefei 230601, PR China
[b] Data Science and Cloud Service Research Center, Shanghai University of Finance and Economics, Shanghai, 200133, PR China
[c] Department of Communication Engineering, North University of China, Taiyuan, Shan'xi 030051, PR China

ARTICLE INFO

Article history:
Received 27 September 2016
Received in revised form 12 January 2017
Accepted 26 January 2017
Available online 30 January 2017
Communicated by C.R. Doering

Keywords:
Complex networks
Multiple spreaders
Heuristic clustering algorithm

ABSTRACT

The problem of influence maximization in social networks has attracted much attention. However, traditional centrality indices are suitable for the case where a single spreader is chosen as the spreading source. Many times, spreading process is initiated by *simultaneously* choosing multiple nodes as the spreading sources. In this situation, choosing the top ranked nodes as multiple spreaders is not an optimal strategy, since the chosen nodes are not sufficiently scattered in networks. Therefore, one ideal situation for multiple spreaders case is that the spreaders themselves are not only influential but also they are dispersively distributed in networks, but it is difficult to meet the two conditions together. In this paper, we propose a heuristic clustering (HC) algorithm based on the similarity index to classify nodes into different clusters, and finally the center nodes in clusters are chosen as the multiple spreaders. HC algorithm not only ensures that the multiple spreaders are dispersively distributed in networks but also avoids the selected nodes to be very "negligible". Compared with the traditional methods, our experimental results on synthetic and real networks indicate that the performance of HC method on influence maximization is more significant.

1. Introduction

Many social, technological and biological systems can be described in terms of networks where nodes represent the elements of the systems and edges define the possible interaction patterns among nodes. The roles of nodes in social networks are often distinct, how to design effective algorithms to identify influential nodes in social networks is related to maintaining the global functionality of the system, developing efficient strategies to control epidemic spreading, accelerating information diffusion, promoting new products, and so on [1–8].

So far, lots of centrality indices have been proposed to identify influential spreaders in networks, such as degree centrality [9], betweenness centrality [10], closeness centrality [11]. Kitsak et al. proposed a k-shell decomposition method to identify the most influential spreaders, and which is better than degree centrality in many real networks [1]. But it tends to assign many nodes that have different spreading capability to the same k-shell value. In particular, which assigns all nodes of tree-like networks to 1-shell. Thus, some methods were proposed to overcome the low resolution of k-shell [6,12–14]. In addition, Radicchi et al. have shown that the nonbacktracking centrality is a highly reliable metric to identify top influential spreaders in social networks [15].

Most of the above mentioned methods mainly focus on how to find "top influential spreaders", that is to say, if one node is chosen as a *single* spreader origin, which *one* should be chosen to maximize the spreading coverage. In this case, proposed indices only need to consider the influence of node itself, but do not consider the interaction effects from other nodes. We call this situation is a single spreader case. However, it is usually that a set of different nodes are *simultaneously* chosen as spreading sources in many spreading processes, such as rumors, opinions, advertisements, and so on. Therefore, the identification of multiple influential spreaders in complex networks is also of theoretical and practical significance, however, this problem has not been well solved. An intuitive way of choosing multiple spreaders is that top ranked nodes who are sorted based on a centrality index (e.g., degree centrality, betweenness centrality, and so on) are selected. However, it may be not the optimal strategy since the top ranked nodes tend to have large overlap in their spreading process, leading to the redundancy of spreading [1,2]. For multiple spreaders case, an effective method should not only consider the influence of nodes themselves but

[*] 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.physleta.2017.01.043
0375-9601/© 2017 Elsevier B.V. All rights reserved.

also consider the dispersibility, therefore, the problem of how to identify multiple influential spreaders is more intricate. Kempe et al. have proved that the issue of finding the most multiple influential spreaders is a NP-hard optimization problem [16]. Recently, some attempts have been made along this line. For instance, Morone et al. offered a framework for the set of optimal influencers in networks by mapping the problem onto optimal percolation problem. However, the method detects the influential nodes one by one rather than *simultaneously*. Namely, the node with the highest value of CI is firstly removed, then the values of CI for remaining nodes should be recalculated, which surely increases the computation complexity [17]. Zhao et al. have proposed a method to obtain the effective multiple spreaders by generalizing the idea of the graph coloring problem to complex networks [18]. Since the distance between the multiple spreaders are not far away in sometimes, the method was improved in Ref. [19]. In the networks with community structures, Hu et al. found that the nodes with the largest degree in each community have the good performance on spreading promotion, and selecting the hub node in each community as multiple influential spreaders [20] is a good choice, but the number of communities may hinder the effectiveness of the method. Thus, how to identify the multiple influential spreaders in social networks is still an important and challenging problem [21–24].

The guiding ideology of selecting multiple spreaders is that the distance among multiple spreaders is relatively scattered, and spreaders themselves are also important. But it is almost impossible to meet both conditions together, we only try to find a tradeoff between them. In this paper, a heuristic clustering (HC) algorithm is proposed to obtain the multiple influential spreaders. In this algorithm, nodes are classified into different clusters based on one similarity index, where the number of clusters equals to the number of multiple spreaders. And the center nodes in clusters are selected as the multiple influential spreaders when the heuristic clustering process is finally stable. Experimental results in synthetic networks and real networks indicate that HC algorithm not only guarantees the selected spreaders are sufficiently scattered but also avoids to be "insignificant". Therefore, the performance of HC algorithm on influence maximization is better than other centrality indices.

The layout of the paper is as follows: Firstly, the descriptions of our method are presented in Sec. 2, and several typical centrality indices, SIR epidemic model and data sets are also introduced in this section. Then the experimental results are presented in Sec. 3. Finally, conclusions are summarized in Sec. 4.

2. Materials and methods

2.1. Heuristic clustering algorithm

An undirected and un-weighted network is represented by $G = (N, M)$ with N nodes and M 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.

If the number of multiple spreaders is m, the details of the heuristic cluster algorithm based on similarity index are the followings.

Step 1: Define similarity matrix. Because the clustering process is implemented based on the similarity between pair of nodes, a similarity matrix is defined at first. There are many ways to define the similarity, in this paper, we use the well-known local path (LP) similarity index in link prediction to define the similarity matrix, since such a similarity index provide a good tradeoff of accuracy and computational complexity [25,26]:

$$S = A^2 + \lambda \cdot A^3, \tag{1}$$

where $0 < \lambda < 1$ is a free parameter, small value of λ means less influence of long paths. $(A^2)_{xy}$ is the number of common neighbors of nodes x and y, which is also equal to the number of different paths with length 2 connecting x and y, and $(A^3)_{xy}$ is the number of different paths with length 3 connecting x and y. Hence, our HC method is a semi-local method. In our paper, we mainly set $\lambda = 0.5$, and we also check the effect of the value of λ in Figs. 5, 6 and 7.

Step 2: Form different clusters. We first randomly select m nodes as the initial centers to cluster nodes, denoted by $D = \{v_1, v_2, \cdots, v_m\}$. For each node $v_k \bar{\in} D$, the similarity $S_{v_k v_i}$ between node v_k and $v_i \in D$, $i = 1, \cdots, m$ is calculated according to Eq. (1), if there is a node $v_i \in D$ such that $S_{v_k v_i}$ is maximum, and then assign node v_k to a cluster whose center is v_i. Therefore, all nodes are classified into m clusters, denoted by $C_1, C_2, \cdots, C_m$;

Step 3: Update center of each cluster. For cluster C_t, $t = 1, \cdots, m$, according to the similarity matrix S, define the significance of node $v_x \in C_t$ in the cluster C_t as $B(x) = \sum_{v_y \in C_t} S_{v_x v_y}$, then select the node with the highest value of significance as the new center of each cluster, that is to say, the set D is updated;

Step 4: Select multiple spreaders. Repeat Step 2 and Step 3 until the algorithm is convergent. At last, the nodes in the set D are viewed as the m multiple influential spreaders.

2.2. Centrality indices

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

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

$$DC(i) = \sum_{j=1}^{N} a_{ij}. \tag{2}$$

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

$$BC(i) = \sum_{s \neq i \neq l} n_{sl}^{i} / n_{sl}, \tag{3}$$

where n_{sl} and n_{sl}^{i} are the number of shortest paths between nodes s and l, and the number of shortest paths between s and l that pass through node i, respectively.

The k-shell (KS) decomposition method is implemented by the following steps: Firstly, one-degree nodes are removed and keep deleting the existing one-degree nodes until all nodes' degrees are larger than one. All of these removed nodes are 1-shell. Then remove the two-degree nodes and keep deleting until all nodes' degrees are larger than two, and include them into 2-shell. This procedure continues until all nodes have been assigned to a k-shell [1,27].

By generalizing the graph coloring in complex network, the degree coloring (DCC) method was proposed in Ref. [18,19] to identify multiple influential spreaders, which can be summarized as follows: **1)** sort the nodes in a descending order according to their degrees, such that $k(1) \geq k(2) \geq, \cdots, \geq k(N)$; **2)** define a color function π to color each node i with a color m, i.e., $\pi(i) = m$, initially, let $\pi(1) = 1$; **3)** let $C(m) = \{i | \pi(i) = m\}$, where $C(m)$ is a set containing nodes with the same color label m. If an uncolored node j is not connected to the nodes in $C(m)$, then $\pi(j) = m$; **4)** let $m := m + 1$, then choose a node at the top positions of the ranking list from the uncolored node set and back to step 3. The process ends once all the nodes are colored. Finally, the m-top