PHYSA: 17144 Model 3Gsc pp. 1–10 (col. fig: NIL)

ARTICLE IN PRESS

Physica A xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Physica A

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

Q1 Effective identification of multiple influential spreaders by DegreePunishment

Q2 Xiaojie Wang [a], Yanyuan Su [c], Chengli Zhao [a], Dongyun Yi [b,a,*]

[a] College of Science, National University of Defense Technology, Changsha, 410073, China

[b] State Key Laboratory of High Performance Computing, National University of Defense Technology, Changsha, 410073, China

[c] School of Economics and Management, Yanshan University, Qinhuangdao, 066004, China

HIGHLIGHTS

- We propose a method to select multiple influential spreaders in networks.
- The proposed method introduces a punishing strategy to do the selections sequentially.
- Low complexity ensures the feasibility and validity of our method in large-scale networks.

ARTICLE INFO

Article history:
Received 14 October 2015
Received in revised form 12 February 2016
Available online xxxx

Keywords:
Spreading influence
SIR model
Heuristic method

ABSTRACT

With the rapid development of social networks, how to effectively identify a small group of nodes to maximize their spreading influence becomes a crucial topic. Traditional centrality-based methods are often very simple but not so effective compared to other complex methods. In this paper, we propose a heuristic method to select spreaders sequentially by carrying out a punishing strategy to the neighbors of those already selected spreaders. We use the Susceptible–Infected–Recovered (SIR) model to evaluate the performance by considering the number of infected nodes in the end. Experiments on four real networks show that our method outperforms traditional centrality-based methods and several heuristic ones.

1. Introduction

Nowadays, social networks [1–5] are changing the world in almost every aspects. From online making friends to offline dating, they have facilitated our lives a lot. In the area of social networks, measuring the spreading influence of nodes is a crucial topic, which brings out a problem of Influence Maximization (IM) [6–8]. IM is a problem about how to effectively select a small group of nodes that can maximize the spreading influence. Though lots of work have been done to measure the importance of single node [9–14], an efficient method to select a bunch of influential spreaders simultaneously is still lacking.

Generally, the researches about IM can be divided into three categories. (1) Centrality-based methods consider basic centrality indices of networks. For instance, *Degree* centrality suggests that nodes with higher degree are more important than others [15]. *Betweenness* [16,17] and *Closeness*, [18] focus on the geometric location of nodes, which are often meaningful but time-consuming. *PageRank* [19] is a well-known random-walk based centrality. In Ref. [20], Kitsak et al.

* Corresponding author at: State Key Laboratory of High Performance Computing, National University of Defense Technology, Changsha, 410073, China.
E-mail address: dongyun.yi@gmail.com (D. Yi).

http://dx.doi.org/10.1016/j.physa.2016.05.020
0378-4371/© 2016 Published by Elsevier B.V.

proposed the *k-shell* centrality, which used the *k*-shell decomposition to evaluate the influence of nodes. In Ref. [21], Zhao et al. proposed a combination method and got good results. (2) Heuristic methods select the spreaders one by one according to some special heuristic criterions. *DegreeDistance* [22] is a simple one, which restricts the minimal distance among the selected nodes. *SingleDiscount* [23] and *DegreeDiscount*, [23] are designed for Independent Cascading (IC) model. The basic idea is treating the selected nodes and normal ones separately. (3) Greedy methods [24–27] use brute force (or with some optimizations) to select spreaders. These kinds of methods can get the best performance among all methods, while high computational complexity often makes them unavailable for large-scale networks.

In this paper, we propose a heuristic method named *DegreePunishment* as a tradeoff between the simple degree centrality and other complex methods. The method selects the nodes one by one. When a node is selected, a punishment strategy is carried out to its neighbors via paths with a weaken factor. To evaluate the performance, we use the SIR model [28] to measure the spreading influence of nodes selected by the proposed method together with several other ones. Compared with several centrality-based methods and heuristic ones, our method performs best or close to the best with very low computational complexity. Moreover, we investigate the relations between our method and others.

Following sections are organized as follows. In Section 2, we briefly revisit several centrality-based methods and heuristic ones, and propose our method at the end. In Section 3, we use the SIR model and evaluate the performance in a toy network. Section 4 is all about numerical analysis. We introduce the four datasets and compare the performance of all the methods at first, and then discuss the relations between our method and other degree-related ones. Conclusions are given in Section 5.

2. Measures for spreading influence

Consider an undirected network $G(V, E)$, where V is the set of nodes and E is the set of edges. Assuming that N denotes the number of nodes, M denotes the number of links, and $A = (A_{ij})_{N\times N}$ denotes the adjacency matrix of the network. Our aim is to choose l nodes among all to maximize their spreading influence in the network.

In this paper, we consider nine algorithms to achieve our goal: *Degree*, *Betweenness*, *Closeness*, *PageRank*, *k-shell*, *DegreeDistance*, *SingleDiscount*, *DegreeDiscount*, and finally our *DegreePunishment*. The former five are centrality-based methods, and the latter four are heuristic ones.

2.1. Centrality-based methods

Many centrality-based methods [29] have been proposed to rank nodes. In the rankings, the top ranked nodes are considered more influential, while bottom-ranked nodes are normal. Thus it is naive to choose *top* $-\, l$ ranked nodes among all and assume they can maximize the spreading influence in the network.

Degree is maybe the simplest method among all centrality-based ones. A node with higher degree is likely to influence more nodes than a node with lower degree, which is a common natural in the real world. Taking social networks as an example, a person with a huge number of fans is often famous, who can attract many persons believing his words, transmitting his message, or following his style. However, many recent studies argue that higher degree does not always correspond to higher influence, while lots of things may count [20].

Betweenness [16,17] measures the extent to which a node locates on shortest paths between nodes pair. For a specific node, its betweenness is defined as the fraction of shortest paths between all nodes pairs that pass though it. In message-passing scenario, the betweenness of a node can serve as a guide to measure its influence over the flow of information between others [30]. Especially in the Internet, busy hubs often have extremely high betweenness. For the nodes with the highest betweenness, we also call them bottlenecks [31], intermediaries [32], or structural holes [33]. The betweenness of node v is denoted by $C_B(v)$,

$$C_B(v) = \sum_{s\neq v\neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{1}$$

where σ_{st} denotes the number of shortest paths between nodes s and t, and $\sigma_{st}(v)$ denotes the number of shortest paths that pass through v.

Closeness [18] focuses on the geometry location of a node v in the network. Originally, *Closeness* is defined as the reciprocal of mean distances between a node and others. However, that kind of definition cannot be used in disconnected networks [34,35]. In Ref. [34], Dangalchev gives a more robust definition.

$$C_C(v) = \sum_{t\neq v} \frac{1}{2^{d(v,t)}} \tag{2}$$

where $d(v, t)$ denotes the distance between nodes v and t.

PageRank [19] is a random walk based ranking method which also serves as a core algorithm in the famous Google search engine. In *PageRank*, every node v is given a score $PR(v)$ to quantify the probability of being browsed. The score $PR(v)$

Please cite this article in press as: X. Wang, et al., Effective identification of multiple influential spreaders by DegreePunishment, Physica A (2016), http://dx.doi.org/10.1016/j.physa.2016.05.020