Physica A 436 (2015) 833–845

Contents lists available at ScienceDirect

Physica A

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

Improving detection of influential nodes in complex networks

Amir Sheikhahmadi [a,1], Mohammad Ali Nematbakhsh [a,*], Arman Shokrollahi [b,c,1]

[a] *Department of Computer Engineering, University of Isfahan, Isfahan, Iran*
[b] *Department of Mathematics and Computer Science, West Virginia University, Morgantown, WV 26506, USA*
[c] *Center for Artificial Intelligence and Network Science, Sanandaj, Kurdistan, Iran*

HIGHLIGHTS

- The goal is to identify the most influential nodes in complex networks.
- We propose DegreeDistance and improve it in two phases, FIDD and SIDD.
- We take into account distance of seeds as well as the influence score.
- We investigate the rate of unique nodes influenced by our methods.
- The SIDD outperforms other measures by choosing a more appropriate seed set.

ARTICLE INFO

Article history:
Received 17 December 2014
Received in revised form 29 March 2015
Available online 5 May 2015

Keywords:
Centrality measure
DegreeDistance centrality
IC model
Influential nodes
Complex networks

ABSTRACT

Recently an increasing amount of research is devoted to the question of how the most influential nodes (seeds) can be found effectively in a complex network. There are a number of measures proposed for this purpose, for instance, high-degree centrality measure reflects the importance of the network topology and has a reasonable runtime performance to find a set of nodes with highest degree, but they do not have a satisfactory dissemination potentiality in the network due to having many common neighbors ($\mathrm{CN}^{(1)}$) and common neighbors of neighbors ($\mathrm{CN}^{(2)}$). This flaw holds in other measures as well. In this paper, we compare high-degree centrality measure with other well-known measures using ten datasets in order to find a proportion for the common seeds in the seed sets obtained by them. We, thereof, propose an improved high-degree centrality measure (named *DegreeDistance*) and improve it to enhance accuracy in two phases, FIDD and SIDD, by put a threshold on the number of common neighbors of already-selected seed nodes and a non-seed node which is under investigation to be selected as a seed as well as considering the influence score of seed nodes directly or through their common neighbors over the non-seed node. To evaluate the accuracy and runtime performance of DegreeDistance, FIDD, and SIDD, they are applied to eight large-scale networks and it finally turns out that SIDD dramatically outperforms other well-known measures and evinces comparatively more accurate performance in identifying the most influential nodes.

---

[*] Corresponding author.
E-mail address: nematbakhsh@eng.ui.ac.ir (M.A. Nematbakhsh).

[1] These authors equally contributed to this work.

http://dx.doi.org/10.1016/j.physa.2015.04.035
0378-4371/

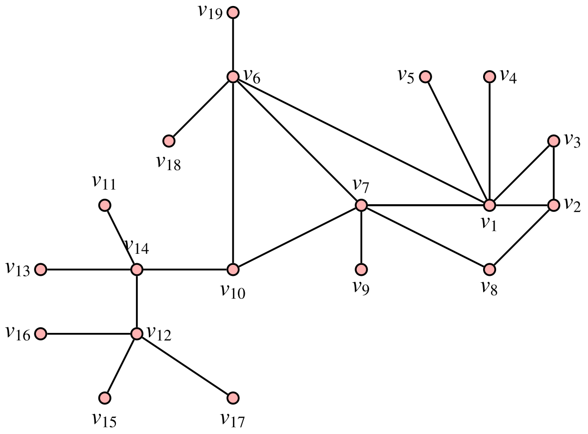

Fig. 1. A sample network which demonstrates that we get better propagation if the seed nodes (v_1 and v_{14}) are chosen in an appropriate distance of each other.

1. Introduction

Identifying the most influential nodes is a pivotal challenge and is of high importance due to its efficacious applications in complex networks, such as proliferation or ceasing the influence over social and economic networks or giving publicity to a product, organization, or venture [1–4], prevention and control of infectious diseases, understanding the function of the human brain and mental disorders [5,6], ranking web pages properly in search engines results [7,8], further analysis of the most enriched processes in biological systems and therapeutic targets [9]. Typically in social networks where the number of users is considerably increasing, one of the goals is maximizing or minimizing the spread of influence through influential nodes. The compulsive, entertaining environments of these networks and the wide diversity of services these systems provide, are making them a proper place for amusement, training, propaganda, etc. [10]. Everyday, we see a huge amount of goods and products advertisements, campaign people ads, and, etc. over these networks. Accepting an advertisement by a user and sharing it with friends and again friends with their friends actively publicize it and facilitates propagation [11–13]. It basically takes advantage of users to advertise products without too much sustained efforts rather than direct interaction which is very costly. On the other hand, the result of this process may be more efficient if friends have confidence in one another [14–18]. This interactive marketing technique is known as "viral marketing" which induces social networking services and other technologies to pass along a marketing message by finding and convincing the most influential individuals [11–17,19,20]. Shortly after, some immediate questions come up like what is the influential node? and how can they be identified? Indeed it is not practically feasible to select all these typical nodes to start propagation due to a shortage of funds and time-consuming, expensive process. Accordingly, the problem is to find an optimal subset of nodes within the network that are able to spread the influence and information as efficient and effective as possible. Previous literature address the maximization problem as "maximizing the spread of influence" [21,22].

Any complex network can be modeled as a directed or undirected network (or graph) consisting of nodes (vertices) and links (edges). Due to conspicuous lack of information about nodes in some complex networks (e.g. social networks), a fairly large amount of scientific studies have considered the structural parameters [23–26,18,27,28]. Then, nodes have been ranked based on the topology of the network and the location of each node in the network. In these approaches, nodes have been evaluated based on measures such as high-degree (or simply degree), betweenness, closeness, etc., and those with the highest/lowest measure have been taken as influential nodes (seeds) to start any desired propagation activities over the network. In this paper, we first scrutinize these measures and figure out a rate of intersection of the seed sets obtained by these measures. Another noteworthy observation is that if seeds in these seed sets are not identical, they are very close to one another so that they are either neighbors or neighbors of neighbors of each other. So, we perceive that the neighborhood overlapping of seeds of different seed sets obtained by these measures is prominent. Hence, these seed sets influence almost the same collection of nodes in the network. Fig. 1 displays a small network and, as we can see, nodes v_1, v_2, v_6, v_7 show high-degree centrality which are adjacent to each other, however by choosing v_1 and v_{14} which are in an appropriate distance of each other, we can achieve a more effective propagation.

Hereinafter, we use the following concepts and notations throughout the paper: The *distance* between two nodes v and w, denoted by $d(v, w)$, is the length of a shortest path between them. We say that a node w is an ith *neighbor* ($i \in \mathbb{Z}^+$) of nodes $v_1, v_2, \ldots, v_r$, $r \geq 1$, if $d(v_1, w) = d(v_2, w) = \cdots = d(v_r, w) = i$. Let $\mathrm{N}^{(i)}(v_1, v_2, \ldots, v_r)$ denote the family of all ith neighbors of nodes $v_1, v_2, \ldots, v_r$, and $\mathrm{N}^{(i)}$ if nodes are not specified. If $A = \{v_1, v_2, \ldots, v_r\}$, we use the short notation $\mathrm{N}^{(i)}(A)$. In some network science and graph theory texts, $\mathrm{N}^{(1)}(v)$ and $\mathrm{N}^{(2)}(v)$ are referred to as neighbors of v and neighbors of neighbors (second order contiguity) of v, respectively. A node z is said to be an ith *common neighbor* of nodes $v_1, v_2, \ldots, v_r$, $r \geq 1$, if $z \in \bigcap_{h=1}^{r} \mathrm{N}^{(i)}(v_h)$. We denote the set of all ith common neighbors of nodes $v_1, v_2, \ldots, v_r$ by $\mathrm{CN}^{(i)}(v_1, v_2, \ldots, v_r)$, and $\mathrm{CN}^{(i)}$ if v_h's ($h = 1, 2, \ldots, r$) are not specified. We define $\mathrm{CN}^{(1,2)} = \mathrm{CN}^{(1)} \cup \mathrm{CN}^{(2)}$. A node w is said to be in *distance threshold*, d_{td},