Physica A 433 (2015) 344–355

Contents lists available at ScienceDirect

Physica A

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

Using global diversity and local topology features to identify influential network spreaders

Yu-Hsiang Fu [a], Chung-Yuan Huang [b,*], Chuen-Tsai Sun [a]

[a] Department of Computer Science, National Chiao Tung University, 1001 Ta Hsueh Road, Hsinchu 300, Taiwan

[b] Department of Computer Science and Information Engineering, School of Electrical and Computer Engineering, College of Engineering, Chang Gung University, 259 Wen Hwa 1st Road, Taoyuan 333, Taiwan

HIGHLIGHTS

- We combine global diversity and local features to identify influential nodes.
- A robust and reliable two-step framework is presented as a node ranking measure.
- Results from a series of experiments indicate our method performs well and stably.

ARTICLE INFO

Article history:
Received 3 June 2014
Received in revised form 30 January 2015
Available online 16 April 2015

Keywords:
Node diversity
Entropy
Social network analysis
k-shell decomposition
SIR epidemic model

ABSTRACT

Identifying the most influential individuals spreading ideas, information, or infectious diseases is a topic receiving significant attention from network researchers, since such identification can assist or hinder information dissemination, product exposure, and contagious disease detection. Hub nodes, high betweenness nodes, high closeness nodes, and high *k*-shell nodes have been identified as good initial spreaders. However, few efforts have been made to use node diversity within network structures to measure spreading ability. The two-step framework described in this paper uses a robust and reliable measure that combines global diversity and local features to identify the most influential network nodes. Results from a series of Susceptible–Infected–Recovered (SIR) epidemic simulations indicate that our proposed method performs well and stably in single initial spreader scenarios associated with various complex network datasets.

1. Introduction

The network-spreading phenomenon is the focus of studies ranging from information diffusion via online social media sites, to viral marketing, to epidemic disease identification and control, to cascading failures in electrical power grids and the Internet, among many others [1–9]. Strategies for identifying key spreaders are being established and tested to accelerate information dissemination, increase product exposure, detect contagious disease outbreaks, and execute early intervention strategies [10]. Topological structure is a core concept in this network spreading identification process [1,2].

In social network analyses, centrality measures for identifying influential network nodes are broadly categorized as local or global [3,7,11]. Degree centrality, defined as the number of nodes that a focal node is connected to, measures node involvement in a network. However, techniques favored by most researchers for measuring the influence of network nodes fail to consider the importance of global topological structures. The two most widely used global centrality measures for

* Corresponding author. Tel.: +886 3 211 8800x3474; fax: +886 3 211 8700.
E-mail address: gscott@mail.cgu.edu.tw (C.-Y. Huang).

http://dx.doi.org/10.1016/j.physa.2015.03.042

overcoming these limitations are betweenness and closeness. Betweenness centrality, which assesses the degree to which a node lies on the shortest path between two other nodes, determines network flow. Closeness centrality is defined as the inverse sum of the shortest distances from a focal node to all other nodes. Influence is tied to the occupation of advantageous network positions. Three basic sources of advantages are high degree, high closeness, and high betweenness. In simple network structures, these advantages tend to vary individually. In complex networks, the potential exists for significant disjunctures among these position characteristics, meaning that a spreader's location may be advantageous in some ways and disadvantageous in others.

In addition to centrality measures, results from a k-shell decomposition analysis indicate that network nodes located in core layers are capable of spreading throughout a much broader area than nodes located in peripheral layers [1,2]. Although the spreading capability of each node differs, those with similar k-shell values are perceived as having equal importance. A method for ranking the network spreading ability of nodes in terms of degree centrality in identical k-shell layers for purposes of adjusting rank lists has been proposed [8]. To rank spreaders, a method referred to as mixed degree decomposition (MDD) adds otherwise ignored degree nodes to the decomposition process [3,6,12]. Still, researchers have shown a tendency to overlook the importance of network topology and node diversity, despite their positive correlations with factors such as community economic development [13]. Further, the entropy values of locations visited by users are positively correlated with the numbers of social ties those same users have in a social network [14]. The combined entropy values of node degree, betweenness, and closeness centralities have been applied to create complex network visualizations [15,16].

Inspired by past studies of network topology and node diversity, we used the entropy concept to develop a robust and reliable method for measuring the spreading capability of nodes, and for identifying super-spreader nodes in complex networks. This measure can be used to analyze the numbers of global network topological layers and local neighborhood nodes that are affected by specific individual nodes. Our assumption is that k-shell decomposition [1,2] can be used for purposes of global analysis, with nodes having high degrees of global diversity and local centrality capable of penetrating multiple global layers and influencing large numbers of neighbors in the local layers of a complex network. To measure node influence, we propose a two-step framework for acquiring global and local node information within complex networks. In the first step, global node information is obtained using algorithms such as a community detection algorithm for complex networks [5,17,18] or a k-shell decomposition algorithm for core/periphery network layers, after which entropy is used to evaluate the global diversity of network nodes. In the second step, local node information is acquired through the use of various types of local centrality, including degree centrality. Last, global diversity and local features are combined to determine node influence. In our experiments, spreading ability was measured as the total number of recovered nodes over time. We compared the spreading capability of the proposed measure and the local/global centralities of the social network using an SIR (susceptible–infective–recovered) epidemic simulation [2,19,20] with various social network datasets [21–25].

2. Background

To represent a complex social network, let an undirected graph $G = (V, E)$, where V is the node set and E the edge set of the network. Let $n = |V|$ indicate the number of network nodes and $m = |E|$ the number of edges. Network structure is represented as an adjacency matrix $A = \{a_{ij}\}$ and $a_{ij} \in R^n$, where $a_{ij} = 1$ if a link exists between nodes i and j, otherwise $a_{ij} = 0$.

Degree (or local) centrality is a simple yet effective method for measuring node influence in a complex network. Let $C_d(i)$ denote the degree centrality of node i. A high-degree centrality indicates a large number of connections between a node and its neighbors. $NB_h(i)$ denotes the set of neighbors of node i at a h-hop distance. The degree centrality of node i is therefore defined as

$$C_d(i) = |NB_h(i)| = \sum_{j=1}^{n} a_{ij} \tag{1}$$

where $|NB_h(i)|$ is the number of neighbors of node i at the h-hop distance; in most cases, $h = 1$ [7].

Betweenness centrality or dependency measures the proportion of the shortest paths going through a node in a complex network. Let $C_b(i)$ denote the betweenness centrality of node i. A high betweenness value indicates that a complex network node is located along an important communication path. Accordingly, the betweenness centrality of node i is defined as

$$C_b(i) = \sum_{s \neq t \neq v \in V} \frac{Q_{st}(i)}{Q_{st}} \tag{2}$$

where $Q_{st}(i)$ is the number of shortest paths from node s to node t through node i, and Q_{st} the total number of shortest paths from node s to node t [3,7,11].

Closeness (also known as global) centrality measures the average length of the shortest paths from one node to other nodes. Let $C_l(i)$ denote the closeness centrality of node i. A high closeness centrality value indicates that a node is located in