ARTICLE IN PRESS

Journal of Network and Computer Applications xx (xxxx) xxxx–xxxx

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

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

Using epidemic betweenness to measure the influence of users in complex networks

Sheng Wen[a,b,*], Jiaojiao Jiang[a,b], Bo Liu[a,b], Yang Xiang[a,b], Wanlei Zhou[a,b]

[a] Centre of Cyber Security Research, Deakin University, Australia
[b] Deakin University, School of Information Technology, 225 Burwood Highway, Burwood, VIC 3125, Australia

ARTICLE INFO

Keywords:
Epidemics
Betweenness
Centrality measures
Network security

ABSTRACT

Betweenness is a measure of the centrality of a node in a network, and is normally calculated as the fraction of shortest paths, random walk paths or flow units between node pairs that pass through the node of interest. Betweenness is, in some sense, a measure of the influence a node possesses over the spread of information in the network. However, the traditional betweenness is based on the information dissemination from one node to another. This is conceptually not suitable for the epidemics in which information is disseminated from one node to multiple neighboring nodes and destinations. To address this problem, we propose a novel betweenness measure based on epidemics. The epidemic betweenness counts the average number of the following nodes influenced by the node of interest after it becomes the epidemic source or an intermediary. This measure reflects the potential influence of a node to any epidemic in complex networks. To justify this measure, we introduce real complex networks and estimate the average influential scale of each node in epidemics through a large number of simulations. We compare the simulation results to those of the epidemic betweenness and another seven classic measures, such as Eigenvector and Katz. We further provide correlation studies to expose the differences of the epidemic betweenness in capturing influential nodes. We find that the epidemic betweenness is exclusively the measure that accurately present the potential influence of each node in epidemics. Finally, as an example of application, the epidemic betweenness measure explains the finding in recent research that unpopular users (nodes with small degree) could also lead to large cascades of epidemics.

1. Introduction

Epidemic has long been a critical problem in various forms of networks. For example, rumors spread incredibly fast in online social networks, such as Facebook and Twitter (Doerr et al., 2012). Computer viruses spread throughout the Internet and compromise millions of computers (Wang et al., 2014). In Smart Grids, isolated failures lead to the rolling blackouts in cities (Yan et al., 2013). Influential users can initiate and conduct the dissemination of information more efficiently than normal users. Therefore, influential users in networks are normally more responsible for large cascades of epidemics.

Researchers have developed many methods to expose influential users in networks. The simplest measure is the degree of node which counts the number of edges incident on a node (Sabidussi, 1966). Generally, the large-degree nodes correspond to the popular users in social networks. The eigenvector centrality (Bonacich, 1972) is an extension of the degree measure. Unlike the degree which weights every neighboring nodes equally, the eigenvector centrality weights the neighboring nodes according to their importance. The Katz centrality (Katz, 1953) is another extension of the degree measure. The node degree stands for the number of direct neighbors, while the Katz centrality counts the number of all reachable nodes and the contributions of distant nodes are penalized. A more sophisticated centrality measure is closeness (Freeman, 1979), which is the mean geodesic (i.e., shortest-path) distance from the node of interest to all other reachable nodes. The closeness measures the efficiency of a node distributing information to any node in networks.

Another important class of centrality measures are betweenness measures. In 1977, Freeman (1977) proposed the *shortest-path betweenness* which is defined as the fraction of shortest paths between node pairs in a network that pass through the node of interest. The shortest-path betweenness is the simplest and most widely used betweenness measure, which is usually regarded as a measure of influence a user possesses over information spreading between any pair of users. However, in most networks, information does not spread only along the geodesic paths. To address this problem, in 1991,

* Corresponding author at: Deakin University, School of Information Technology, 225 Burwood Highway, Burwood, VIC 3125, Australia
E-mail addresses: wesheng@deakin.edu.au (S. Wen), jjiaoj@deakin.edu.au (J. Jiang), bo.liu@deakin.edu.au (B. Liu), yang@deakin.edu.au (Y. Xiang), wanlei@deakin.edu.au (W. Zhou).

http://dx.doi.org/10.1016/j.jnca.2016.10.018
Received 6 November 2015; Received in revised form 3 October 2016; Accepted 24 October 2016
Available online xxxx
1084-8045/

Freeman et al. (1991) proposed a more complex betweenness measure, usually known as the *flow betweenness*. The flow betweenness is based on the idea of maximum flow, which is defined as the number of flow units through the node of interest when the maximum flow is transmitted between node pairs. For these two betweenness measures, the information needs to "know" the ideal route (shortest or maximum-flow path) from one node to another. However, the ideal routes between node pairs are normally unknown during the transmission, and the information wanders around randomly in the network until it reaches the destination. Accordingly, in 2005, Newman (2005) proposed a new betweenness measure based on random walks. The *random-walk betweenness* counts how often a node is traversed by a random walk between node pairs.

1.1. Motivation

In this paper, we study the betweenness measures and their application of capturing influential nodes in the epidemics. The traditional betweenness measures have proved of great value to the analysis of node influence in complex networks (Comin and da Fontoura Costa, 2011; Newman, 2010; Wen et al., 2014). However, these measures focus on the dissemination of information from one node to another (e.g. through shortest paths or random-walk paths). This is conceptually not suitable for the epidemics in which the information is sent from one node to multiple neighboring nodes and destinations (i.e. through multi-paths to all reachable nodes). According to our investigation, the nodes exposed by traditional betweenness measures in the one-to-one communication may be different from those which can lead to large-scale epidemics in the one-to-multi scenarios. Therefore, we are motivated to propose a new betweenness measure based on epidemics in complex networks.

We restrict the context of the epidemic influence and epidemic betweenness. Information in epidemics goes from the source to the neighboring nodes and then to the neighbors of neighboring nodes until all reachable nodes receive the information. As epidemic incidents may start from any nodes in networks, the node of interest may be the epidemic source or an intermediary forwarding the receiving to the neighboring nodes. As for the influence of a node, we normally consider a node to be influential to the epidemics if this node can influence a large number of following nodes after it becomes influenced by the epidemics. We then give the formal definition of the epidemic betweenness as follows:

Epidemic betweenness: the epidemic betweenness of an arbitrary node i, $b_{EP}(i)$, is the expected number of nodes that are influenced directly or indirectly by node i after i becomes influenced by epidemics. The value $b_{EP}(i)$ is averaged by the epidemic incidents that start from all possible sources in the network.

1.2. Contributions

The **primary** contribution of this paper is the novel epidemic betweenness measure which fixes the drawbacks of traditional betweenness measures. This new measure reflects the potential influence of a node to any epidemic in a complex network.

To study the significance of the epidemic betweenness, we conduct a series of experiments in various synthetic and real complex networks. First, we estimate the influence of each node through a large number of simulation runs. The estimated influence is compared to the epidemic betweenness of each node. The results suggest that the new measure can *accurately* present the epidemic influence of each node in complex network. Second, we record the sets of influential nodes exposed by the epidemic betweenness measure and various classic centrality measures, and derive their intersections to the simulation results. As the influential nodes exposed by the epidemic betweenness measure are much closer to the nodes indicated by the simulations than all the other measures, the epidemic betweenness is the *exclusive* measure to present the influence of nodes in complex networks. We finally investigate the correlations between the epidemic betweenness measure and the classic measures. The results disclose that the traditional shortest-path and random-walk betweenness measures are the two which are moderately correlated to the newly proposed epidemic betweenness measure.

The rest of this paper is organized as follows. In Section 2, we use examples to explain the necessity of the epidemic betweenness. We present the computation of the epidemic betweenness in Section 3. The evaluation and analysis are presented in Sections 4 and 5. Section 6 presents an example of its applications, followed by the related work and conclusion in Sections 7 and 8.

2. Problem statement

The traditional betweenness measures (i.e., shortest-path (Freeman, 1977), flow (Freeman et al., 1991) and random-walk (Newman, 2005) betweenness) have long been employed to locate the influential nodes in complex networks. However, these measures are conceptually not suitable for the epidemics in which information spreads from one node to multiple receivers rather than the transmission from one to another. In this section, we discuss the reason why the newly proposed epidemic betweenness measure is necessary and important. For the convenience of readers, we list the major notations used in this paper in Table 1.

2.1. Epidemic vs. shortest-path

The shortest-path betweenness centrality is defined as the fraction of the geodesic (i.e., shortest) paths between node pairs that pass through the node of interest in a network. To be precise, suppose that $g_i^{(st)}$ is the number of geodesic paths from node s to t that pass through node i, and $g^{(st)}$ is the total number of geodesic paths from s to t. Then the shortest-path betweenness centrality of node i is

$$b_{SP}(i) = \frac{\sum_{s<t} g_i^{(st)}/g^{(st)}}{(1/2)(n-1)(n-2)}, \tag{1}$$

where n is the total number of nodes in the network. The shortest-path betweenness stands for the ability of a node relaying information between an arbitrary pair of nodes in the network under the consideration of always choosing the shortest paths.

The problem: In epidemics, the geodesic paths that have the same distance between a pair of nodes may provide different propagation probabilities. The nodes on these paths will have the same shortest-path betweenness. However, as the information proceeds their epidemic distributions differently in the paths with different propagation probabilities, these nodes may have different epidemic influences. Therefore, the shortest-path betweenness is not suitable for exposing the influence of nodes in epidemics.

Table 1
Major Notations Used in This Paper.

Notation	Explanation
$b_{SP}(i)$	The shortest-path betweenness of an arbitrary node i.
$b_{FL}(i)$	The flow betweenness of an arbitrary node i.
$b_{RW}(i)$	The random-walk betweenness of an arbitrary node i.
$X_i(t)$	The state of user i at discrete time t.
$S(t)$	The number of susceptible users at time t.
$I(t)$	The number of infected nodes at time t.
$R(i, t)$	The probability of user i not receiving or accepting the information.
$v(i, t)$	The probability of user i becoming contagious.
$E(A_{i\|s})$	The overall influence of node i in the epidemic incident $A_{i\|s}$.
$E(A_{i\|s}^t)$	The influence of node i at time t.
$\delta_{ij}{}^t$	The ratio of node i's contribution to the infection of node j at time t.
$b_{EP}(i)$	The epidemic betweenness of an arbitrary node i.