Physica A 461 (2016) 840–853

Contents lists available at ScienceDirect

Physica A

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

Using a two-phase evolutionary framework to select multiple network spreaders based on community structure

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, Chang Gung University, 259 Wen Hwa 1st Road, Taoyuan 333, Taiwan

HIGHLIGHTS

- Network community structures are used to identify multiple network spreaders.
- A two-phase evolutionary framework is proposed for finding community structures.
- Our framework produced satisfactory community quality without performance loss.
- A community center measure for selecting network spreaders is described.
- Distinct community structure benefits multi-network spreader dissemination.

ARTICLE INFO

Article history:
Received 26 January 2016
Received in revised form 18 May 2016
Available online 16 June 2016

Keywords:
Genetic algorithm
Community detection
Network spreading
Social network analysis
Multiple network spreaders

ABSTRACT

Using network community structures to identify multiple influential spreaders is an appropriate method for analyzing the dissemination of information, ideas and infectious diseases. For example, data on spreaders selected from groups of customers who make similar purchases may be used to advertise products and to optimize limited resource allocation. Other examples include community detection approaches aimed at identifying structures and groups in social or complex networks. However, determining the number of communities in a network remains a challenge. In this paper we describe our proposal for a two-phase evolutionary framework (TPEF) for determining community numbers and maximizing community modularity. Lancichinetti–Fortunato–Radicchi benchmark networks were used to test our proposed method and to analyze execution time, community structure quality, convergence, and the network spreading effect. Results indicate that our proposed TPEF generates satisfactory levels of community quality and convergence. They also suggest a need for an index, mechanism or sampling technique to determine whether a community detection approach should be used for selecting multiple network spreaders.

1. Introduction

Influential network spreader identification can be used to accelerate or hinder the spread of information, ideas, or diseases in social and complex networks [1–5]. Strategies tied to identified spreaders can be used for tasks such as increasing product exposure ranges in marketing, monitoring contagious disease outbreaks, designing and executing disease

* Corresponding author.
E-mail addresses: yuhsiangfu.cs98g@nctu.edu.tw (Y.-H. Fu), gscott@mail.cgu.edu.tw (C.-Y. Huang), ctsun@nctu.edu.tw (C.-T. Sun).

http://dx.doi.org/10.1016/j.physa.2016.06.042
0378-4371/

intervention strategies [6], and speeding up Internet information dissemination or diffusion [2,4,5]. Thus, researchers are investing considerable time and resources into formulating influential network spreader identification schemes.

In some social network analyses, centrality measures and network structures are used to evaluate node importance. These measures can be classified as *local* or *global* [7–9]. Degree centrality is considered a simple and effective tool for measuring node importance: nodes with high degrees of centrality (e.g., hub nodes with multiple connections) have been shown to exert greater network influence [7–9]. In contrast, global centralities can be measured in terms of betweenness, closeness, and k-shell decomposition [1,3,7–9]. High betweenness indicates the location of a node on many of the shortest communication paths; high closeness indicates that a node serves as a network center with shorter average lengths of the shortest paths to other nodes in the network. High k-shell values (determined using k-shell decomposition) indicate node location within core network layers. A diffusion-based approach such as the PageRank (PR) algorithm can also be used to measure node importance in a network [10]. High PR value indicates node connections with many important neighbors. One research team has identified a method for measuring global diversity and local features that they describe as robust and less sensitive to the effects of network spreading [7,8].

However, centrality methods are insufficient for identifying multiple network spreaders, and are therefore not considered appropriate solutions for many practical marketing and disease prevention applications. Since community structure has been identified as an important network property along with small-world [11], scale-free [12], and fractal properties [13,14], community detection approaches are increasingly being used to identify network community structures or social groups to select multiple network spreaders [9,15–21]. Borgatti has suggested the use of two methods to identify important network nodes for purposes of defining a key-players problem (KPP) for selecting multiple network spreaders: KPP-Pos, a set of key players that are maximally connected to all others nodes, and KPP-Neg, a set of key players that are removed, thereby producing a residual network that has the lowest possible cohesion [22].

According to Borgatti's KPP-Pos definition, finding a set (or simply a limited number) of key players can be reduced to a set/vertex cover problem or a 0/1 Knapsack problem. For successful community detection, the primary challenge is identifying a network's community structure, which in turn supports the identification of multiple network spreaders in communities that are considered representative. Accordingly, a network community detection problem can be reduced to a graph partition problem, but the reduced problem is NP-Complete (NPC) [23]. Evolutionary computing approaches such as genetic algorithms (GAs) [24–29], ant colony optimization (ACO) [30], and particle swarm optimization (PSO) [31] are suitable for finding approximate solutions to NPC problems. GAs have at least two advantages: approximate solutions can be efficiently identified according to specific limitations, and the content of approximate solutions can be analyzed, understood and preserved—that is, community structure information can be extracted from chromosomes and stored in files.

Borgatti's KPP-Neg definition is connected with the problem of finding a set of key players via mapping onto an optimal percolation problem, which minimizes many-body system energy [32]. In optimal percolation, a collective influence (CI) algorithm is used to identify a minimal (optimal) set of key players (influences) that, if removed, could break down an entire network into small, disconnected components at a certain threshold. Weak nodes with small numbers of connections, surrounded by hierarchical hub coronas, emerge from identified optimal sets of influencers in both theoretical and real-world networks. If weak nodes that are capable of spreading information throughout a network are identified as activated or immune, that increases the potential for reducing the number of vaccines that must be distributed and used during a large-scale pandemic. The use of weak nodes is also considered better than centrality-based methods for identifying approximate optimal critical thresholds.

In the present study, network community structures are used to select K network spreaders (or K key players) based on a combination of the KPP-Pos definition, resource limitations (i.e., the number of network spreaders), and community detection. The main criterion for identifying community structures is maximizing modularity [9,15–17,21]. Further, the primary condition for selecting network spreaders is that the nodes must have the highest numbers of intra-community connections and lowest numbers of inter-community connections [22]. According to Kitsak [3], a secondary condition is that distances between spreaders must be considered to determine the extent of network spreading overlap. Selected network spreaders are expected to spread information, ideas, or viruses from inside to outside communities in a network. However, since determining the K number of communities is a difficult task, K usually remains unknown. We therefore propose a *two-phase evolutionary framework* (TPEF) as a trade-off between performance efficiency and community compactness for automatically determining K. In phase one, an appropriate K value is automatically determined based on network topology sampling. In phase two the focus is on optimizing the phase one network partition. Multiple network spreaders are chosen from the identified communities.

For our preliminary experiment, we used the LFR benchmark model [33,34] to perform tests involving a partition-based straightforward GA (SGA) [29], a locus-based GA (LGA) [35–37], and our proposed TPEF to compare execution times, community structure quality, and solution convergence. Results indicate that TPEF is capable of determining an appropriate number of communities (i.e., K number) and generating satisfactory community structures. According to our network spreading simulation results, in the distinct community structure case (i.e., $u < 0.2$) we found that using a community detection approach to select K network spreaders resulted in a much wider spreading range than the centrality-based methods. However, in cases of indistinct community structures (i.e., $u \geq 0.2$), good performance was noted when centrality-based methods were used to select K network spreaders, with results similar to those produced by the community detection approach. In other words, there is at least one benefit to applying community detection methods to assist with K network spreader selection in situations marked by distinct community structures, otherwise centrality-based methods can be used