Social Networks 54 (2018) 41–49

Contents lists available at ScienceDirect

Social Networks

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

A new scalable leader-community detection approach for community detection in social networks

Sara Ahajjam*, Mohamed El Haddad, Hassan Badir

Laboratory of Information and Communication Technologies, National School of Applied Sciences, ENSA, Tangier, Morocco

ARTICLE INFO

Article history:

Keywords:
Leader
Community detection
Big graph
Centrality
Social network
Similarity
Big data
Graph theory

ABSTRACT

Studying social influence in networks is crucial to understand how behavior spreads. An interesting number of theories were elaborated to analyze how innovations and trends get adopted. The traditional view assumes that a minority of members in a society possess qualities that make them exceptionally persuasive in spreading ideas to others. These exceptional individuals drive trends on behalf of the majority of ordinary people. They are loosely described as being informed, respected, and well connected. The leaders or influential are responsible for the dissemination of information and the propagation of influence. In this paper, we propose a new scalable and a deterministic approach for the detection of communities using leaders nodes named Leader-Community Detection Approach LCDA. The proposed approach has two main steps. The first step is the leaders' retrieval. The second step is the community detection using similarity between nodes. Our algorithms provide good results compared to ground truth membership community.

1. Introduction

Complex networks represent complex systems in different areas. They can be modeled by a graph, where nodes represent the actors of the system, connected by edges to describe different types of relationships. Discovering communities is a fundamental problem in network science, which has attracted vast attention in recent years (Xie et al., 2013; de Arruda et al., 2014). Several research studies have addressed the problem of community detection. Community detection aims to find clusters as sub-graphs within a given network (Social Network Analysis, 2017), with the purpose of finding the communities using the information embedded in the network topology. A community is defined as a set of nodes highly interconnected, and loosely connected to other nodes in the network.

Within certain communities, some nodes play more important roles in diffusion of information, ideas, and innovation within those communities. They are the catalysts of influence. Therefore, it has motivated many researchers to look for an efficient method to find the most influential members in social networks. For example, it is in trading companies and banks interest to find active and influential parties in their existing network and potentially extend their network to include other parties (Wang et al., 2011a).

Identifying influential or leaders nodes in networks can be regarded as ranking important nodes, and it has become one of the main problems in network-based information retrieval and mining (Domingos and Richardson, 2001). In epidemic spreading, it is vital to find the important nodes to understand the dynamic processes, which could shed some light on immunizing modular networks. In biological systems, key nodes are identified in the communities for the purpose of treatment, for example in the case of lung cancer, the treatment option entails destroying cancer cells while protecting normal cells (Domingos and Richardson, 2001).

Identification of influential nodes relies on quantitative characterization of the node in terms of their importance to community structure, centrality in particular. Centrality aims to identify the most important actors in a social network, which determines the social influence and power of each actor within such network. Centrality can be local or global (Gao et al., 2014; Networks, 2010; Renoust, 2014). The local methods, which include degree and betweenness centrality, use the local features of the node to determine its importance. Degree centrality measures node involvement in a network and is determined by the number of nodes that a focal node is connected to. The betweenness centrality assesses centrality in a network of nodes based on shortest paths.

However, most network node researchers fail to consider global topological structures. Closeness centrality could be used to over-

* Corresponding author.
E-mail addresses: ahajjamsara@gmail.com (S. Ahajjam), elhaddad.mohamed@gmail.com (M. El Haddad), hbadir@gmail.com (H. Badir).

https://doi.org/10.1016/j.socnet.2017.11.004
0378-8733/

come such limitation, and it is assessed based on inverse of the sum of the shortest distances between each node and every other node in the network. Also, K-shell decomposition analysis shows that network nodes in core layers are capable of spreading to broader areas compared to those in peripheral layers. Since these central nodes have the ability to diffuse their influence to the whole network faster than the rest of the nodes, they would be regarded as the most influential spreaders.

This paper deals with two important research subjects in computer science: the community detection and the leader detection in complex networks. The majority of the proposed algorithms detect the community first followed by identification of the leader of a given community. The main limitation associated with such methodology is that prior knowledge of the number and the size of the final partitions are required. In this paper, a scalable and deterministic approach is presented to detect communities in social networks using leaders' nodes. In this approach, leader nodes of the networks that are responsible for the dissemination of influence are detected, then communities will be formed around the leaders using similarity between nodes, i.e. using different edge based similarity measures. In Section 2, relevant work and studies will be reviewed to provide border perspective and insight to the proposed subject. In Section 3, fundamentals of proposed algorithms will be presented in detail. In Section 4, experimental results will be presented, and the final section offers concluding remarks and further application.

2. Preliminaries & related works

Leader detection and community detection in complex networks, particularly in social networks have been the subject for wide research studies in recent years. The Centrality and prestige measures are used to determine the importance of a node in undirected and directed networks respectively, where prestigious nodes, the influential or the leader nodes are identified. Prestige and centrality provide reliable analysis of the nature of relationship and connection between nodes, which is important to understand a wide range of phenomena.

In computer-science literature, several methods have been applied to analyze user influence in social network and community leader detection in online social networks. The leader detection approaches are divided into two main groups: global and local methods. The global methods emphasize on all the network topology (betweenness centrality), while the local methods focus on local positions, i.e. individual nodes (degree centrality). In the context of social science, the influence can be defined as bargaining power, control over information, or level of persuasiveness (Khorasgani et al., 2010). Khorasgani et al. suggested a new approach to detect leader nodes that consider outliers, i.e. nodes that are not associated with any leader. This algorithm is inspired by K-means algorithm. In K-means algorithm, k nodes will be selected randomly, and other nodes will be assembled at their closest leaders to form communities. For each community, new leader will be identified gathering other new followers until no node moves. For each community, the centrality of each member is calculated and the node with the highest degree will be appointed as the new leader (Kernighan and Lin, 1970). Another approach of Bae et al. used the coreness centrality to estimate the spreading influence of a node. The coreness centrality is estimated by the k-shell indexes of the adjacent neighbors of the node. Therefore, the coreness centrality of a node is presented as the sum of the k-shell values of its adjacent nodes (Bae and Kim, 2014).

The existing methodologies used for community detection can be divided into two main categories, i.e. graph partitioning and classification. The major drawback of partitioning of graphs is that prior knowledge of the number of groups to be detected (Newman, 2013; Bader et al., 2013; Fortunato, 2011) is required. For example, in the study carried out by Kernighan and Lin, the algorithm of leader nodes detection in complex networks is based on the partitioning of graphs. This algorithm tries to find a section of the graph minimizing the number of edges between partitions by trading vertices between these partitions, and results are generated by introducing the size of each partition (Fortunato, 2011), and result can vary significantly given the assumption of size and number of partitions of the graphs.

Classification methodology is also used by researchers to analyze data and partition based on a measure of similarity between partitions. Approaches based on a partitioned classification require prior knowledge of the number of communities to be detected (Fortunato, 2011; Yang et al., 2014; Wu et al., 2013). The major problem to overcome is to select an appropriate distance for better data classification (Yang et al., 2014), which explains why classification methods are generally more appropriate for networks with hierarchical structure. Agglomerative methods and divisive methods are two main groups of methods used in hierarchical clustering. The agglomerative methods attempt to recursively merge small communities into larger communities based on a measure of proximity between communities. It is a bottom up approach, each node will be defined as a cluster, and clusters will be combined at each step, until pairs of clusters are merged as one moves up the hierarchy. The divisive methods attempt to identify the inter-links and remove them to gradually isolate communities, i.e. they initially put all nodes in a complete graph and they remove the links between nodes with the lowest similarity (de Arruda et al., 2014). Some other studies utilize the spectral classification; in the approach proposed by Yang et al., the community is defined based on three properties: community structure to define strong and weak communities, community membership to detect members of the communities, and overlapping property, which considers the number of connections between the node and the corresponding community. The result obtained by these methods will heavily depend on the choice of the similarity measure used initially (Yang et al., 2014).

Some other research studies propose methods that combine the two predefined subjects: leader detection and community detection. In the Leader-Follower algorithm, internal structure of a community should be defined, i.e. the community should be a clique and is formed with a leader and at least one "loyal follower", with loyal follower defined as a node that only has neighbors within a single community. The leader is a node whose distance is less than at least one of its neighbors. Nodes will be allocated to the community in which a majority of their neighbors belongs by destroying the links arbitrarily. However, removing parasite communities will lead to loss of information (Wu et al., 2013).

Qin et al. proposed a novel centrality guided clustering for detecting leader nodes and communities by choosing the vertex with the highest centrality as a starting point. The approach is carried out with three steps: grouping will cluster vertices in G into different groups, merging groups with a large percentage of overlap, and contract those vertices in the same groups to form a new vertex (Zhou et al., 2014). Zhang et al. proposed a greedy algorithm based on user preferences (GAUP) to operate the top-k influential users, based on the model Extended Independent Cascade. During each cycle i, the algorithm adds a record in the selected set such that the vertex S with the current set S maximizes propagation of the influence. This means that the vertex selected in round i is the one that maximizes the incremental propagation influence in this cycle. This algorithm calculates the user's preferences for different subjects, and combines traditional greedy algorithms and preferences calculated by LSI user and calculates an approximate solution of the problem of maximizing the influence of a specific topic. This