Computers and Electrical Engineering 000 (2018) 1–10

Contents lists available at ScienceDirect

Computers and Electrical Engineering

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

Overlapping community detection using superior seed set selection in social networks☆

Belfin R.V.[a,*], Grace Mary Kanaga E.[a], Piotr Bródka[b]

[a] Department of Computer Sciences Technology, Karunya Institute of Technology and Sciences, Coimbatore, India
[b] Department of Computational Intelligence, Faculty of Computer Science and Management, Wrocław University of Science and Technology, Wrocław, Poland

ARTICLE INFO

Article history:
Received 21 June 2017
Revised 2 March 2018
Accepted 7 March 2018
Available online xxx

Keywords:
Seed set
Centrality
Social network
Overlapping community discovery
Seed-centric

ABSTRACT

Community discovery in the social network is one of the tremendously expanding areas which earn interest among researchers for past one decade. There are many already existing algorithms. However, new seed-based algorithms establish an emerging drift in this area. The basic idea behind these strategies is to identify exceptional nodes in the given network, called seeds, around which communities can be located. This paper proposes a blended strategy for locating suitable superior seed set by applying various centrality measures and using them to find overlapping communities. The examination of the algorithm has been performed regarding the goodness of the identified communities with the help of intra-cluster density and inter-cluster density. Finally, the runtime of the proposed algorithm has been compared with the existing community detection algorithms showing remarkable improvement.

1. Introduction

Some active social networks are very vibrant and add a huge amount of users at a particular interval. Each user in the social network will have their specifications and affinity in using certain products. Consider a company which wants its new products to be promoted to the customers through the social network. One way of promoting their products is to broadcast the product information to everyone in the network. This form of information propagation will be expensive and might be treated as spam by most of the social network users. Most of the victorious promotions in the social networks will be conducted by liking and sharing the content with the imminent ones in the network. The uncertainty arises when the marketing team wants to know " Where to start the promotion?". This situation prompts the researchers to work on an algorithm for finding the seed node from which the dispersion of information starts. For the larger network, there exists a necessity of having more seed nodes to proclaim the information faster across the whole network. This seed selection strategy also can be implemented in community detection to find out hidden communities available in the social network. The communities found will be crucial for the marketing team in a company to target marketing.

Heterogeneous networks are commonly utilized for modeling intercommunications in real-world systems in several areas, such as sociology, biology, knowledge spreading and transferring and numerous other areas [1]. One key topological property of real-world heterogeneous networks is that nodes are organized in tightly affiliated groups that are loosely connected to

☆ Reviews processed and recommended for publication to the Editor-in-Chief by Guest Editor Dr. D. Peter.

* Corresponding author.
E-mail address: belfin@karunya.edu (B. R.V.).

https://doi.org/10.1016/j.compeleceng.2018.03.012
0045-7906/

each other. Such groups are called as communities.Nodes forming a community are commonly agreed to share common proprieties and get involved in the same kind of functions [2].

There are many community detection methods proposed by researchers for the past few years. Disjoined community discovery is one of the biggest group of community detection method where every node in the network belongs only to one community [3]. The second group is overlapping community detection methods where one node can be a part of more than one community. Overlapping communities are valid in real social networks like Facebook, Twitter, and LinkedIn because the users will be involved in various ventures and associate with various communities [4]. Another aspect of the problem with overlapping community detection is the fuzzy nature of nodes concerning the degree of affinity towards each community in the network [5].

The basic idea of the strategy proposed in this paper is to find out a fraction of superior nodes of the input network, called superior seed set, around which local communities can be computed. The key concepts required to understand the proposed algorithm is given in the preliminaries section.

The rest of the paper is organized as follow. This chapter introduces the reader to basic concepts used in the paper. Chapter 2 presents state of the art in the field of community detection. In the 3rd chapter, the new approach is presented and evaluated in the 4th chapter. All findings are summarized in the last, 5th chapter.

1.1. Preliminaries

1.1.1. Problem definition

Given a graph $G=(V, E)$ with a set of nodes V and a set of edges E, we can denote the graph as an adjacency matrix A such that $A_{ij}=e_{ij}$ where e_{ij} is the edge weight between vertices i and j, or $A_{ij}=0$ if there is no edge. We can also determine various centrality measures C_i for a node i to represent the strength of a node i in the entire graph. Let us assume the graph to be undirected. The conventional methods use any one of the centrality measures concerning the problem to define its seed set $S(G)=\{s_i,s_j,\ldots,s\kappa\}$ where $\{s_i,s_j,\ldots s_k\} \in v$. Some algorithms pick seed set randomly which takes the parameter κ as input, to decide the number of seeds in the set. Every iteration the seeds will be adjusted to take the best possible position to cover the nodes around it. The difficulty in the conventional methods is that the seed set needs to be adjusted according to the circumstances of the problem. The proposed unified model determines the superior seed set S(G) from the centrality measures collectively. This article introduces a threshold value τ, which limits the number of seed nodes selected for the S(G).

1.1.2. Centrality measures

Centrality measures will address the insights concerning a node's status in the whole social network. There are numerous alternatives of centrality measures possible with regard to the nature of the network. The commonly used centrality measures are the degree, betweenness, closeness, and the Eigenvector centrality.

- ***Degree centrality***

The degree is defined as the number of links it has with its neighbors. For an undirected graph, the degree will be the total number of links a node holds. Likewise, for a directed graph the nodes will own both "in-degrees" and "out-degrees". The in-degree of a node is the number of nodes incident on it, whereas, the out-degree of a node is defined as the number of links traveling apart from it.

Let $A=(a_{i,j})$ is being the adjacency matrix of a directed graph. The in-degree centrality x_i of node i is given by:

$$x_i = \sum_k a_{k,i} \tag{1}$$

The out-degree centrality y_i of node i is given by:

$$y_i = \sum_k a_{k,i} \tag{2}$$

- ***Closeness***

Closeness is a measure which was first defined by Freeman in 1978. The real idea behind this measure is to identify the nodes which could influence other nodes in the network faster. The disadvantage of this measure is that it is not applicable to the graph containing separated components.

Suppose $d_{i,j}$ is the length of a geodesic path from i to j, meaning the number of edges along the path. Then the mean geodesic distance for vertex i is:

$$l_i = \frac{1}{n} \sum_j d_{i,j} \tag{3}$$

The mean distance l_i is not like other centrality measures because it gives the least value to the significant central nodes and vice versa. So, to determine closeness the Eq. (3) can be rewritten inversely as:

$$C_i = \frac{1}{l_i} = \frac{n}{\sum_j d_{i,j}} \tag{4}$$