Pattern Recognition Letters 67 (2015) 145–152

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Fuzzy-rough community in social networks ☆

Suman Kundu*, Sankar K. Pal

Center for Soft Computing Research, Indian Statistical Institute, Kolkata 700108, India

ARTICLE INFO

Article history:
Available online 26 February 2015

Keywords:
Social network
Granular computing
Normalized fuzzy mutual information
Community detection
Soft computing
Big Data

ABSTRACT

Community detection in a social network is a well-known problem that has been studied in computer science since early 2000. The algorithms available in the literature mainly follow two strategies, one, which allows a node to be a part of multiple communities with equal membership, and the second considers a disjoint partition of the whole network where a node belongs to only one community. In this paper, we proposed a novel community detection algorithm which identifies fuzzy-rough communities where a node can be a part of many groups with different memberships of their association. The algorithm runs on a new framework of social network representation based on fuzzy granular theory. A new index viz. normalized fuzzy mutual information, to quantify the goodness of detected communities is used. Experimental results on benchmark data show the superiority of the proposed algorithm compared to other well known methods, particularly when the network contains overlapping communities.

1. Introduction

Traditionally, social network is considered to be a theoretical construct useful in social sciences to study relationships between individuals, groups, organizations or even entire society. However, the recent boom in the social network via Facebook, Twitter, WhatsApp, LinkedIn, made it an everyday affair. This provides new research opportunities, especially in Computer Sciences, because the data available from these online social networking sites are dynamic, large scale, diverse and complex. That is, it shows all the characteristics of Big Data such as velocity, volume, and variety.

Since its inception in early 20th century, social networks are represented using graphs [1], and graph analysis has become crucial to understand the features of these networks [2]. Due to the recent revolution in computing (processing) power, one can now handle relatively larger real networks [3] potentially reaching millions of vertices. Accordingly, it leads to a deep change in the way social networks were being analyzed.

Social networks are different from random networks. It shows fascinating patterns, and properties [4]. The degree distribution is skewed, following the power law Barabási [5,6] or truncated geometric distribution [7]. Diameter of the network is found to be very small compare to the size of the network, and the network possesses high concentration of edges in its certain parts forming groups. This last feature, that is, groups with high internal edge density within themselves and low between them characterizes the community structure (or clustering) of the network.

In society, it is possible to find groups, such as families, co-workers' circle, friendship circles, villages, and town that naturally form. Similar to this, in an online social network, we can find virtual groups, which live on the web. For example, in world wide web it will help to optimize the Internet infrastructure [8], in a purchase network it can boost the sell by recommending appropriate products [9], and in computer network it will help to optimize the routing table creation [10]. Again, identifying special actors in the network is also a motivating force behind community detection. For example, central nodes of the clusters, or nodes in the boundary region who act as a bridge between communities, are the special actors who play different important roles within the society.

Therefore, the challenge in community detection is to identify the modules and possibly their hierarchical organization by only using the information encoded in the network topology. Scientists from several disciplines studied the problem for a long time. One of the first studies on community identification was carried out by Rice [11] where clusters are identified in a small political body based on their voting patterns. Later in 1955, Weiss and Jacobson studied community structure within a government agency [12]. They have separated workgroups by removing those people who work with different groups. This idea of removing edges is the basis of several algorithms in recent times [13,14]. Hierarchical [15] and partition based clustering is the more traditional technique to identify communities in a social network where vertices are jointed into groups as per their mutual similarities.

Girvan and Newman [13], presented a new algorithm, aiming at the identification of the edges lying between two communities for

☆ This paper has been recommended for acceptance by G. Sanniti di Baja.

* Corresponding author. Tel.: +91 905 1301 121.
E-mail addresses: suman@sumankundu.info, suman_nsec@yahoo.com (S. Kundu), sankar@isical.ac.in (S.K. Pal).

http://dx.doi.org/10.1016/j.patrec.2015.02.005
0167-8655/

possible removal in order to find the communities. The possible edges were identified based on their centrality values. The concept is considered as the start of modern era in community detection. Since then many new methods have been proposed based on several techniques like label propagation algorithm [16], optimization [17] and Statistical Physics [2]. These involve mainly two strategies for finding communities in a network. The first approach considers a partition of the whole network into disjoint communities (i.e., a node belongs to only one community). The second strategy, on the other hand, allows a node to be a member of multiple communities with equal membership. However, intuitively there could be a third possibility, that is, a node may belong to more than one community with different degrees of associations.

The present article concerns with the third strategy where we propose a novel algorithm for detecting communities, over a new framework of knowledge representation of social networks. This new framework is based on fuzzy granular theory where a granule is constructed around nodes and represented by a fuzzy set. The proposed algorithm takes the set of granules as input and partition them into meaningful communities. After getting all communities we further model them in the framework of rough sets. That is, the nodes surely belonging to a community constitute its lower approximation, and the others possibly belonging to the community are identified as member of "upper - lower" or boundary region. The nodes in boundary region belong to multiple communities with different degrees of association. We assign fuzzy membership for these nodes based on their connectivity with the cores; thereby resulting in unequal membership unlike the previous methods. Therefore, given a social network, the proposed method determines the various communities with fuzzy-rough description defined over a granular model of knowledge representation.

Extended LFR benchmark data [18] is used to test the algorithm and its aspects. In addition to this, we used two real-world benchmark data viz. Zakary Karate Club data [19] and Dolphin Network Data [20] to demonstrate the performance. To quantify the performance, a new index, namely, *normalized fuzzy mutual information* (NFMI) is used. Comparison is made with three well known community detection algorithms of both overlapping and non-overlapping types. Results show superior performance of the proposed method.

The rest of the paper reads as follows: Section 2 contains the proposed fuzzy granular model of the social network and the community detection algorithm along with remarks and notes. Section 3 reports the experimental results and derivation of the new normalized fuzzy mutual information measure. Finally, in Section 4 we conclude.

2. Model and algorithm

2.1. Fuzzy granular model of social network

A social network is viewed as a collection of relationship between actors such as individuals or organization. These actors form macro-level groups with their neighbors, which are often sometime indistinguishable in the process of problem solving. These groups are different as compare to the community or clusters in terms of size and working principles. These are more like closely operative groups exists within a neighborhood. These macro groups resemble the concepts of granules. A granule is considered to be a clump of objects (or points) in the universe of discloser, drawn together by indistinguishablity, similarity, proximity or functionality [21,22].

Again the relationships between nodes, clusters of nodes, interactions between nodes do not lead themselves to precise definition. That is these macro groups have ill-defined boundaries. So, it is appropriate and natural that we represent a social network in the framework of fuzzy granular theory.

A social network presented in fuzzy granular framework is represented by a triple

$\mathcal{S} = (\mathcal{C}, \mathcal{V}, \mathcal{G})$ where

- $\mathcal{V}$ is a finite set of nodes of the network
- $\mathcal{C} \subseteq \mathcal{V}$ is a finite set of granule representatives (1)
- $\mathcal{G}$ is the finite set of all granules around each $c \in \mathcal{C}$

A granule $g \in \mathcal{G}$ around a representative node $c \in \mathcal{C}$ is defined by assigning fuzzy membership values to its neighborhood. When we consider a node's relationship in a social network, the membership value should decrease as distance increases. So, any monotonically non-increasing fuzzy function may represent a granule in a network. Depending upon the network properties and problem in hand one can choose suitable fuzzy function to assign membership values. In our experiments, we use the following fuzzy membership values,

$$\mu_c(v, r) = \begin{cases} 0 & \text{for } d(c, v) > r \\ \dfrac{1}{1 + d(c, v)} & \text{otherwise} \end{cases} \tag{2}$$

Here, $d(c, v)$ is the distance function which indicates a distance from the granule center c to node v in the network. r is considered to be the radius of the granule.

Remark 1. If one wants to capture the maximum information of the network, $\mathcal{C}$ should be equal to $\mathcal{V}$. However, social network data available from online network shows Big Data characteristics. So, a model describing these kinds of networks needs to address the challenging issue of scalability. In this regard, for reducing the execution time of data analysis to a tolerable range one can restrict the number of granules either based on a threshold, set over the cardinality of the granule, or with human intervention.

Remark 2. Distance function $d(c, v)$ can be of any metric depending upon the problem in hand. For example, when we address community detection, one can use

1. the minimum hop distance from node c to v,
2. or, minimum weighted hop distance, i.e. $d(c, v) = \sum_{e \in P} \omega(e)$ where $\omega(e)$ is the weight of the edge e in path P from c to v,
3. or, the reciprocal of the "number of paths" available between c to v in conjunction with the minimum hop distance.

A point to note here is that when social relationships required to be analyzed with non-metric similarity measures for problems such as Homophily or Positional analysis, one may consider a membership function other than Eq. (2) as suited to their problems.

Remark 3. A node of a social network $\mathcal{S}$, can belong to more than one granule and in such scenario, the node will have a different degrees of belongingness to various granules. For a node v having non-zero membership to more than a granule, membership values can be normalized using the following equation such that all these normalized membership values add up to unity.

$$\tilde{\mu_c}(v, r) = \frac{\mu_c(v, r)}{\sum_{i \in \mathcal{C}} \mu_i(v, r)} \text{ such that } \sum_{i \in \mathcal{C}} \tilde{\mu_i}(v, r) = 1. \tag{3}$$

2.2. Fuzzy-rough community detection on fuzzy granular model of social network (FRC-FGSN)

A community is formed when nodes are densely connected, compare to the other parts of the network. In the new knowledge representation scheme of fuzzy granular social network, as stated in Section 2.1, we would like to find out such densely connected groups. The key idea of finding such groups is to identify the granules with dense neighborhood and merge them when they are nearby (merging dense regions). Thus the first step is to find those granules where *granular degree* (Definition 1) exceeds a certain threshold indicating dense region.