Pattern Recognition Letters 000 (2015) 1–8

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 xxx

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,

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

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.

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

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.

Please cite this article as: S. Kundu, S.K. Pal, Fuzzy-rough community in social networks, Pattern Recognition Letters (2015), http://dx.doi.org/10.1016/j.patrec.2015.02.005

Definition 1 (Granular degree)**.** Granular degree of a granule is defined by the cardinality of the fuzzy set representing the granules. So, granular degree of A_c is,

$$\mathcal{D}(A_c) = |A_c| = \sum_{v \in \mathcal{V}} \tilde{\mu_c}(v, r) \tag{4}$$

where r is the radius of the granule. Granular degree is the fuzzy equivalent degree of the node in the center to the granule.

Remark 4 (Crisp equivalence)**.** Let us consider a crisp membership value for the granules. That is, if a node v is connected to the center node c, it will get a membership of 1, and 0 otherwise. Furthermore, consider $r = 1$. Then, the granular degree $\mathcal{D}(A_c) = \sum_{v \in \mathcal{V}} \tilde{\mu_c}(v, 1) = D(c)$, which is nothing but the network degree of node c. $D(c)$ represents the crisp equivalence of granular degree (Eq. (4)).

Definition 2 (θ-Core)**.** A granule A_p is said to be a θ-core with respect to θ, if the granular degree of the A_p is greater or equals to θ, i.e., $\mathcal{D}(A_p) \geq \theta$.

A community may have multiple such θ-cores. The algorithm needs to identify the set of those closeby θ-cores. So the goal is to search for θ-cores which belong to a same community. We call them 'community reachable cores'. In order to find them, let us first define the neighborhood of a granule as in Definition 3.

Definition 3 (Neighborhood of a granule)**.** Neighborhood of a granule A_c is the set of all granules whose centers lies on the support set of A_c, i.e.,

$$\Gamma(A_c) = \{A_i | A_i \in \mathcal{G} \text{ and } i \in \text{Support}(A_c) \forall i \neq c\}$$

where, $\text{Support}(A_c) = \{v | \tilde{\mu_c}(v, r) > 0\}$. r is the radius of the granule.

Based on the neighborhood, thus defined, we can find the θ-cores which are community reachable to each other, i.e., belong to the same community.

Definition 4 (Directly community reachable θ-cores)**.** Granule A_p and A_q are directly community reachable θ-cores, if A_p and A_q are θ-cores and A_p is in the neighborhood of A_q, i.e., if $A_p \in \Gamma(A_q)$ and $\mathcal{D}(A_q), \mathcal{D}(A_p) \geq \theta$.

Definition 5 (Community reachable θ-cores)**.** A granule A_p is community reachable θ-cores to granule A_q if there is a chain of granule centers $p_1, p_2, \ldots, p_n; p_1 = p$ and $p_n = q$ such that $A_{p_{i+1}}$ is directly community reachable θ-cores from A_{p_i}.

Community reachable cores have another notion of connectivity, say, community connected θ-cores, as stated in Definition 6.

Definition 6 (Community connected θ-cores)**.** Two θ-cores A_p and A_q are said to be community connected if there exists a θ-core A_r from which both A_p and A_q are community reachable.

In a network, there might be nodes, which reside at the boundary regions and have neighborhood spread over multiple groups. To represent the notion of this overlapping, we introduce a normalized granular embeddedness measure as in Definition 7.

Definition 7 (Normalized granular embeddedness)**.** For two given granules A_p and A_q, normalized granular embeddedness is defined by the ratio of the cardinality of their intersection and union, i.e.,

$$\mathcal{E}(A_p, A_q) = \frac{|A_p \cap A_q|}{|A_p \cup A_q|}$$

$\mathcal{E} = 0$ implies no overlapping between granules A_p and A_q. $\mathcal{E} = 1$ signifies complete overlapping.

With these new Definitions 1–7, based on fuzzy granular framework of social network, let us define community and orphan nodes of a network as in Definitions 8–10.

Definition 8 (Community)**.** For a given social network $\mathcal{S} = (\mathcal{C}, \mathcal{V}, \mathcal{G}), \theta$, and ϵ, a community $\mathbb{C}$ is a non-empty subset of granules $\mathcal{G}$ satisfying the following conditions:

- $\forall A_p, A_q \in \mathbb{C}$, A_p and A_q are community connected cores
- $\forall A_p \in \mathbb{C}, \mathcal{E}(A_p, \bigcup_{A_q \in \mathbb{C} \setminus A_p} A_q) > \epsilon$

Remark 5. θ may be referred as density co-efficient of the community. That is, as θ increases the more dense communities get detected and in the process the number of orphan nodes (Definition 10) increases. If the parameter θ is chosen too high, it may happen that there is no θ-cores in the system and in such cases the algorithm returns "no community", i.e., all nodes are orphans. This scenario can be avoided by choosing the aforesaid parameter more conservative way, for example, by selecting the mean of granular degrees as θ.

Remark 6. $1/\epsilon$ may be referred as coupling co-efficient of the community. When a higher value of coupling is selected then loosely connected groups get merged into a single community. On the other hand, very low coupling value may result in more communities than desire.

One may note that the communities, thus identified, have fuzzy (ill defined) boundaries. These communities can further be viewed in terms of lower and upper approximations in the framework of rough set theory. That is, each community has a lower approximate region reflecting nodes definitely belonging to, and a boundary (i.e., upper - lower) region reflecting the nodes possibly belonging to. Therefore it may be appropriate to assign fuzzy membership values in (0, 1) to only those nodes which lie within the said (upper - lower) region, and assign unity (1) value to those of lower approximation. The fuzzy-rough communities are accordingly defined (Definition 9).

Definition 9 (Fuzzy-rough community)**.** Let the n communities found for a social network be $\mathbb{C}_1, \mathbb{C}_2, \ldots, \mathbb{C}_n$, and the upper and lower approximation of the ith community be $\overline{\mathbb{C}_i\theta}$ and $\underline{\mathbb{C}_i\theta}$ respectively. Then

$$\begin{aligned} \underline{\mathbb{C}_i\theta} &= \{x | x \in \text{Support}(A_p) \wedge x \notin \text{Support}(A_q); \\ &\qquad \forall A_p \in \mathbb{C}_i \text{ and } A_q \in \mathbb{C}_j; i \neq j\} \\ \overline{\mathbb{C}_i\theta} &= \{x | x \in \text{Support}(A_p); A_p \in \mathbb{C}_i\} \end{aligned} \tag{5}$$

Fuzzy-rough membership function characterizing the community $\mathbb{C}_i$ is defined as,

$$\delta^{\theta}_{\mathbb{C}_i}(x, r) = \begin{cases} 1 & \text{if } x \in \underline{\mathbb{C}_i\theta} \\ \sum\limits_{c \in \underline{\mathbb{C}_i\theta}} \tilde{\mu_c}(x, r) & \text{if } x \in \overline{\mathbb{C}_i\theta} \setminus \underline{\mathbb{C}_i\theta} \\ 0 & \text{Otherwise} \end{cases} \tag{6}$$

where $\tilde{\mu_c}(x, r)$ is defined in Eq. (3).

Definition 10 (Orphans)**.** Let a social network contain $\mathbb{C}_1, \mathbb{C}_2, \ldots, \mathbb{C}_n$ communities. A node p is said to be orphan if $p \notin \overline{\mathbb{C}_i\theta}$ $\forall i$.

Given a social network, the proposed method finds its various communities (Definition 8) with fuzzy-rough description (Eq. (6)) defined over a granular model (Eq. (1)) of knowledge representation. Nodes not being included as a part of any community are designated as orphans. The block diagram of the methodology is shown in Fig. 1 for convenience.

3. Experiment and results

In this section, we evaluate the performance of the proposed algorithm and compare it with other popular community detection algorithms. To compare results, we consider a new index measure, namely, normalized fuzzy mutual information (NFMI).

Please cite this article as: S. Kundu, S.K. Pal, Fuzzy-rough community in social networks, Pattern Recognition Letters (2015), http://dx.doi.org/10.1016/j.patrec.2015.02.005

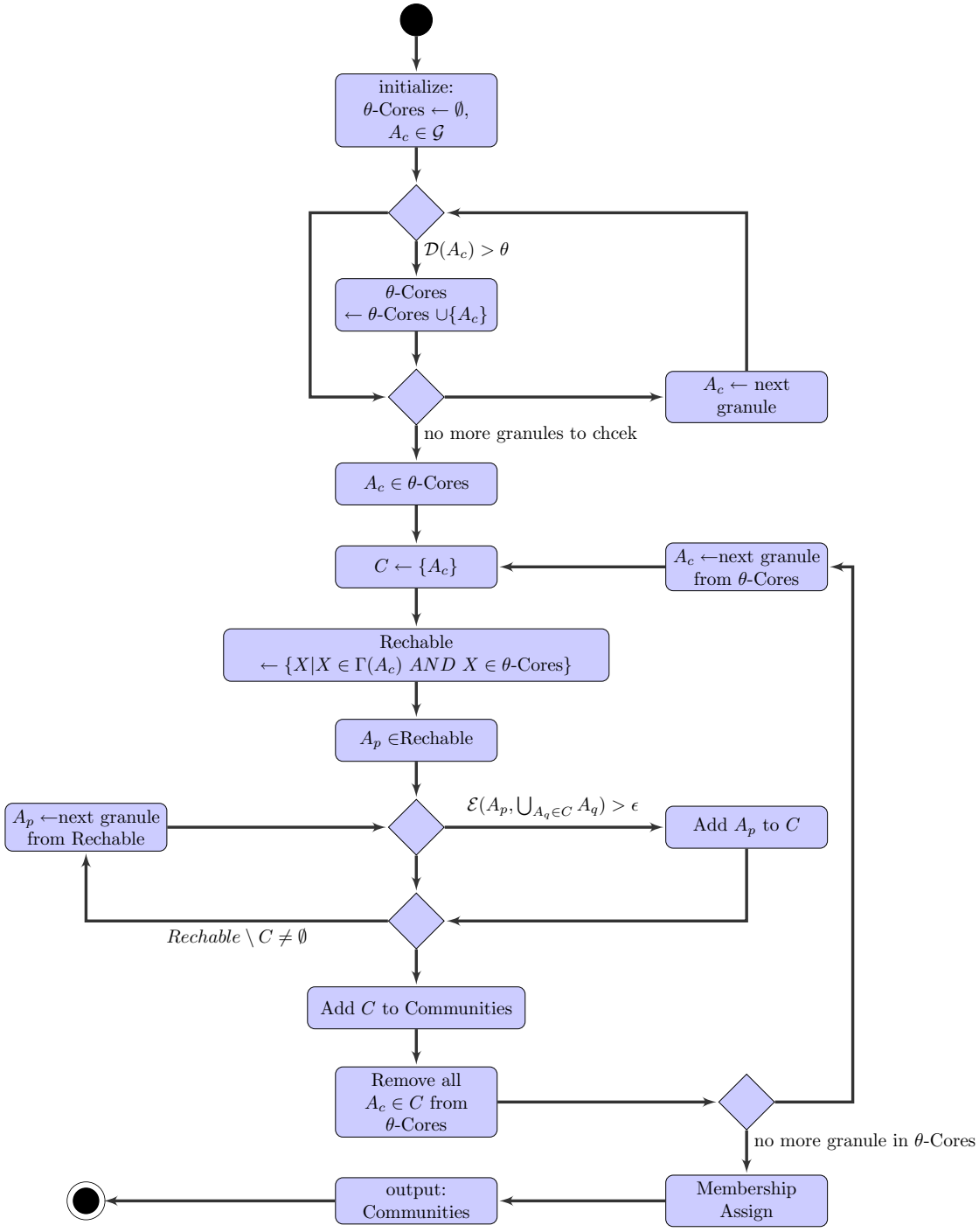


Fig. 1. Block diagram of FRC-FGSN algorithm.

3.1. Normalized fuzzy mutual information

In recent time, a measure based on normalized mutual information [23] has become popular for comparing community structures. However, this measure is suitable for crisp membership values. Fuzzy mutual information, on the other hand, was proposed by Maji and Pal [24] to use in a supervised gene selection algorithm with respect to normal-cancer classification. In case of community detection (which is unsupervised), the numbers of communities is unknown, and the numbers detected by distinct algorithms are also different. Here, we describe a new index measure, namely normalized fuzzy mutual information, suitable for fuzzy community structures.

Let us consider that algorithms X and Y produce two community structures represented by the fuzzy partition matrices $\mathbb{C}^X$ and $\mathbb{C}^Y$. Each row of a partition matrix corresponds to a community. Let the membership that a node v belongs to a community P of $\mathbb{C}^X$ be $m_P^X(v)$. We seek to measure the similarities between $\mathbb{C}^X$ and $\mathbb{C}^Y$. Let there be n nodes in the network. Mutual information of $\mathbb{C}^X$ and $\mathbb{C}^Y$ can be represented as,

$$I(\mathbb{C}^X : \mathbb{C}^Y) = \frac{1}{2}[H(\mathbb{C}^X) - H(\mathbb{C}^X|\mathbb{C}^Y) + H(\mathbb{C}^Y) - H(\mathbb{C}^Y|\mathbb{C}^X)] \quad (7)$$

Please cite this article as: S. Kundu, S.K. Pal, Fuzzy-rough community in social networks, Pattern Recognition Letters (2015), http://dx.doi.org/10.1016/j.patrec.2015.02.005

Here, $H(\mathbb{C}^X)$(or $H(\mathbb{C}^Y)$) is the information contained in $\mathbb{C}^X$(or $\mathbb{C}^Y$) and is defined as:

$$H(\mathbb{C}^X) = -\sum_{P\in\mathbb{C}^X} \lambda_P^X \log_2(\lambda_P^X) \tag{8}$$

where $\lambda_P^X = \sum_i^n m_P^X(i)$ is the fuzzy relative frequency of community $P \in \mathbb{C}^X$.

$H(\mathbb{C}^X|\mathbb{C}^Y)$(or $H(\mathbb{C}^Y|\mathbb{C}^X)$) is the conditional information measure in terms of lack of information of $\mathbb{C}^X$(or $\mathbb{C}^Y$) given $\mathbb{C}^Y$(or $\mathbb{C}^X$). In order to compute the conditional information, we calculate, the joint frequency distribution of two communities P and Q. In an overlapping community structure a node may belong to only P, only Q, both P, Q, or none. Let us now denote these four scenarios respectively as, (i) $P = 1, Q = 0$, (ii) $P = 0, Q = 1$, (iii) $P = 1, Q = 1$ and (iv) $P = 0, Q = 0$. With these notions, the joint frequency distribution of P and Q is as follows

$$\lambda_{00} = \lambda_{(P=0,Q=0)} = \frac{n - |P \cup Q|}{n} = \frac{n - \sum_{i=1}^n \max(m_P^X(i), m_Q^Y(i))}{n} \tag{9}$$

$$\lambda_{01} = \lambda_{(P=0,Q=1)} = \frac{|Q| - |P \cap Q|}{n} = \frac{\sum_i^n m_Q^X(i) - \sum_{i=1}^n \min(m_P^X(i), m_Q^X(i))}{n} \tag{10}$$

$$\lambda_{10} = \lambda_{(P=1,Q=0)} = \frac{|P| - |P \cap Q|}{n} = \frac{\sum_i^n m_P^X(i) - \sum_{i=1}^n \min(m_P^X(i), m_Q^X(i))}{n} \tag{11}$$

$$\lambda_{11} = \lambda_{(P=1,Q=1)} = \frac{|P \cap Q|}{n} = \frac{\sum_{i=1}^n \min(m_P^X(i), m_Q^X(i))}{n} \tag{12}$$

Thus the information measure, in terms of lack of information, is

$$H(P|Q) = H(P, Q) - H(Q) = h(\lambda_{00}) + h(\lambda_{01}) + h(\lambda_{10}) + h(\lambda_{11}) - H(Q) \tag{13}$$

where, $h(x) = -x\log_2(x)$.

We now compute the conditional information measure for a community P, given $\mathbb{C}^Y$, as

$$H(P|\mathbb{C}^Y) = \min_{Q\in\mathbb{C}^Y} H(P|Q); \quad P \in \mathbb{C}^X \tag{14}$$

The conditional information measure of $\mathbb{C}^X$, given $\mathbb{C}^Y$, is then computed as

$$H(\mathbb{C}^X|\mathbb{C}^Y) = \sum_{P\in\mathbb{C}^X} H(P|\mathbb{C}^Y) \tag{15}$$

Similarly, $H(\mathbb{C}^Y|\mathbb{C}^X)$ may be computed.

The normalized fuzzy mutual information is defined as follows:

$$\mathrm{NFMI}(\mathbb{C}^X : \mathbb{C}^Y) = \frac{1}{2}\left[\frac{H(\mathbb{C}^X) - H(\mathbb{C}^X|\mathbb{C}^Y)}{H(\mathbb{C}^X)} + \frac{H(\mathbb{C}^Y) - H(\mathbb{C}^Y|\mathbb{C}^X)}{H(\mathbb{C}^Y)}\right] \tag{16}$$

Higher the value of NFMI larger the similarity (or relevance) between $\mathbb{C}^X$ and $\mathbb{C}^Y$. One may note that higher value may also occur when two communities are nearly complement to each other. In order to avoid this undesirable situation, we enforce the following condition while computing Eq. (13):

$$H(P|Q) = \begin{cases} H(P|Q) & \text{if } h(\lambda_{00}) + h(\lambda_{11}) > h(\lambda_{01}) + h(\lambda_{10}) \\ H(P) & \text{otherwise} \end{cases} \tag{17}$$

Remark 7. The conditional information measure represents the lack of information in a community structure, given another. In an ideal case, i.e., when two comparative community structures are identical, $H(\mathbb{C}^X|\mathbb{C}^Y)$ and $H(\mathbb{C}^Y|\mathbb{C}^X)$ will be zero. Hence, the NFMI($\mathbb{C}^X : \mathbb{C}^Y$) would be equal to 1 (Eq. (16)).

On the other hand, when two community structures are complement to each other, then $H(\mathbb{C}^X|\mathbb{C}^Y)$ and $H(\mathbb{C}^Y|\mathbb{C}^X)$ will be equal to $H(\mathbb{C}^X)$ and $H(\mathbb{C}^Y)$ respectively (Eq. (17)). In this case, the NFMI($\mathbb{C}^X : \mathbb{C}^Y$) score will be 0.

In all the other cases, the value of NFMI($\mathbb{C}^X : \mathbb{C}^Y$) will be between 0 and 1. That means, higher the value of NFMI, more closer the community structures.

To evaluate the performance of two different community detection algorithms, one needs to find the NFMI score of both the output with those of the ground truth. Higher the NFMI value better is the quality of the determined community structures.

3.2. Benchmark

In the evaluation of performance of the proposed method, we used two types of benchmark data. These are synthetically generated networks, and real-world social networks, both with known communities. Description of the data sets is presented below followed by the experimental results.

3.2.1. LFR benchmark

One of the popular benchmark data for comparing community detection algorithms is proposed by Lancichinetti et al. [25] in 2008. It is referred as LFR benchmark data after the name of the authors. Later, it was modified to accommodate more properties of network and communities viz. directed, weighted network and overlapping communities, in Lancichinetti and Fortunato [18]. The idea is to generate network graphs based on few parameters. These parameters include,

- Size of the network N
- Size of the communities (within $C_{\min}$ to $C_{\max}$)
- Mixing parameter, i.e., the average ratio of edges within community and edges with other communities (η)
- Fraction of overlapping nodes (O_n) and
- Number of overlapping communities (O_m)

In the experiments we have fixed the size of the network to 1001 and vary the other variables, and analyze the algorithms and their performance. We compare the proposed algorithms with three popular algorithms. These are, centrality based community detection method proposed by Girvan and Newman [13], Modularity optimization method of Newman [17] and k-clique percolation method (CPM) of Palla et al. [26]. A point to note here is that, CPM can identify overlapping communities whereas the other two comparing methods identify non-overlapping partitions of the network. The benchmark data generated from LFR algorithm for overlapping communities is far from the reality. It considers a fixed number of overlaps for the nodes which is unusual for real world networks. Furthermore, we are assigning different memberships for being in different communities based on the network values, but the generated network assigns unit value to the same. So, it is not the perfect data set to test our algorithms, yet results are convincing, as described below.

First, we vary the η from 0.0 to 1.0 by fixing the fraction of overlap to 0.15 and run all the four algorithms. We measure NFMI of each output with the ground truth. Fig. 2, shows the variation of NFMI with respect to η for these algorithms. As expected, NFMI decreases when η increases in all the cases. For lower values of η, modularity and centrality based algorithms' show better results, but for $\eta \geq 0.3$ the proposed FRC-FGSN shows prominent improvement over all other methods.

In another experiment, we vary the fraction of overlapping nodes (O_n) from 0.0 to 0.5 by fixing the mixing parameter to 0.4. Results

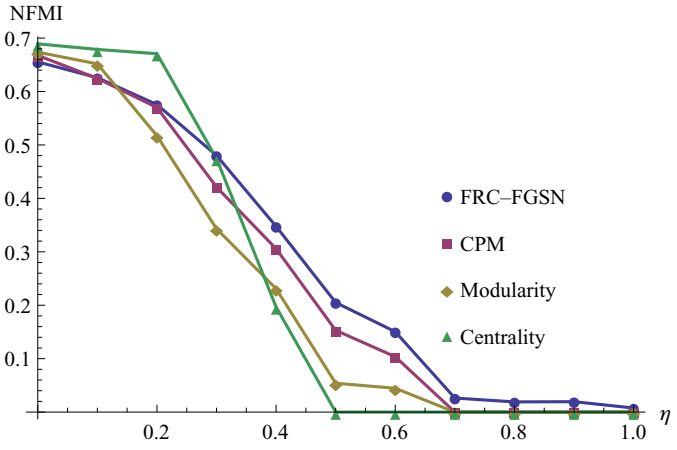


Fig. 2. Comparative results with different values of mixing parameter. Network size: 1001; min community size: 150; max community size: 250.

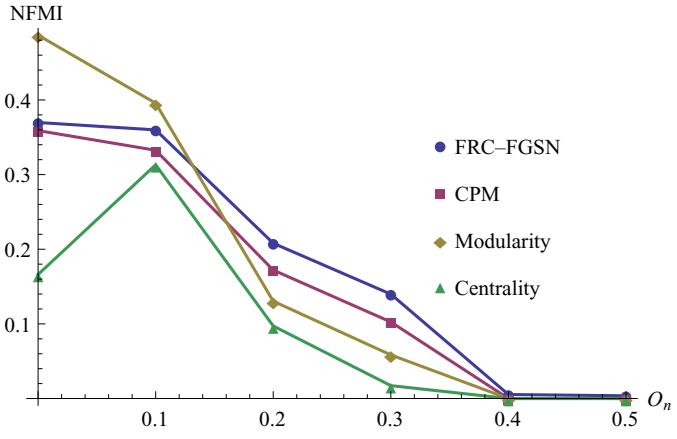


Fig. 3. Comparison showing variation of NFMI for different fraction of overlapping community. Network size: 1001; mixing parameter: 0.4; min community size: 150; max community size: 250.

are reported in Fig. 3. It shows that the proposed FRC-FGSN produces superior performance for O_n ranging from 0.2 to 0.4 and second best for $O_n < 0.2$.

As we mentioned in Remark 1, one may restrict the number of granules to reduce the execution time to a tolerable range. We perform an experiment to observe this phenomenon. The result in this regard, for one of the benchmark networks is shown in Fig. 4. Here, *x*-axis shows the percentage of granules corresponding to the number of nodes in the network. The blue curve with square points shows the

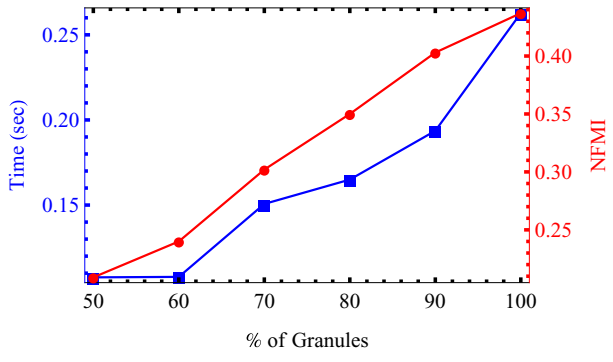


Fig. 4. Plot showing the performance on number of granules for LFR data. (For interpretation of the references to color in the text, the reader is referred to the web version of this article.)

time taken by the proposed FRC-FGSN and the red curve with circular points shows its accuracy in terms of NFMI. As expected, the time and accuracy both decreases as we reduce the number of granules from 100% to 50%. Interestingly the rate of drop in execution time is higher than that of the accuracy. This shows that by reducing the number of granules in a fuzzy granular model of social network one may get execution benefits in the algorithm.

3.2.2. Real world benchmark data

We used two real-world social network data, namely, Zachary Karate Club [19] and Dolphin Social Network [20].

Zachary Karate Club data is shown in Fig. 5(a). This network shows the friendship relations between 34 members of a US Karate Club in 1970s. Fig. 5(b) summarizes the statistics about the data set. The club eventually split into two and the ground truth of the community structure is shown in Fig. 5(c).

The network of frequent associations among 62 bottlenose dolphins living in Doubtful Sound, New Zealand was collected between 1995 and 2001 by Lusseau et al. [20]. The network is an undirected graph of their interactions, and properties of the network are given in Fig. 6.

Although, the ground truth available for these two real world networks does not have any overlapping nodes, yet the results are very promising and close to those of modularity optimization algorithms and better than CPM methods for both the cases. For Karate Club data, it is even much better than the centrality based community detection algorithms. Results of these experiments are shown in Fig. 7.

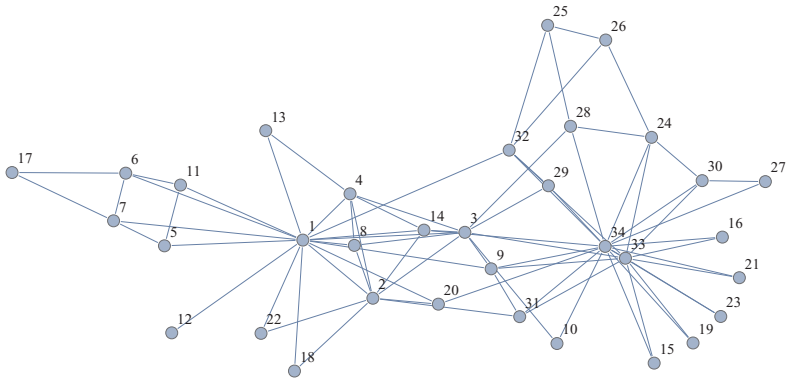
4. Discussions and conclusions

We presented a new algorithm (FRC-FGSN) to identify different communities in a social network. Here, a network is represented by a collection of fuzzy granules. The output communities found are characterized with crisp lower and fuzzy upper memberships, and are designated as “fuzzy-rough communities”. A fuzzy membership is assigned only to those nodes which fall into the boundary (upper - lower) region of a community signifying that a node in that region can belong to multiple communities with different degrees of association. Nodes in the lower approximate region are assigned unity membership reflecting the certainty of the belonging. In the process orphan (nodes with zero membership to all communities) are detected automatically. The proposed framework of knowledge representation is capable of handling uncertainty arising from both fuzziness in boundary and granularity of the community.

In addition to the proposed algorithm, an index, namely, normalized fuzzy mutual information (NFMI) has been defined to quantify the goodness of the identified communities. Larger is the value of NFMI, between two community structures, higher is their similarities. Computation of this measure involves comparison of two fuzzy partition matrices, one corresponding to the identified communities and the other corresponds to those of the ground truth. Here, the best match of each of the identified communities out of those in the ground truth is determined. Amount of matching is quantified in terms of fuzzy mutual information. Normalized aggregate of these matching scores is reflected by the NFMI index, which accordingly quantifies how close the identified communities are to those of the ground truth.

It is shown that the FRC-FGSN algorithm produces superior outcomes as compared to other popularly known community detection algorithms when the network contains overlapped communities. We, reported experimental results conducted with both LFR benchmark data and real-world social network of Zachary Karate Club data and Dolphin Social Network.

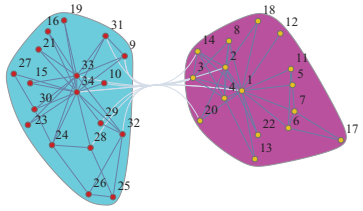
In the proposed knowledge representation scheme, we have considered single relationship between actors. In addition, we considered that the membership of a node in a granule decreases as its distance



(a) Network

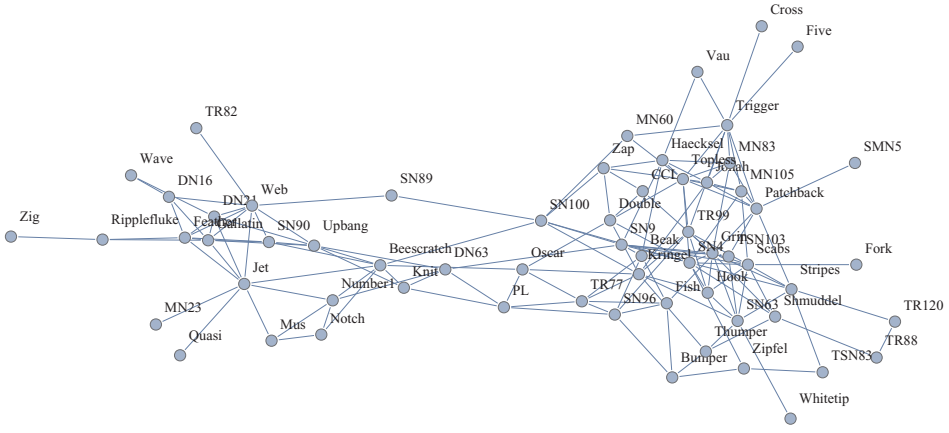
Nodes	34
Edges	78
Nodes in largest Weakly Connected Component	34
Edges in largest Weakly Connected Component	78
Nodes in largest Strongly Connected Component	34
Edges in largest Strongly Connected Component	78
Diameter	5
Avg. clustering coefficient	0.570638

(b) Statistics



(c) Known Community Plots

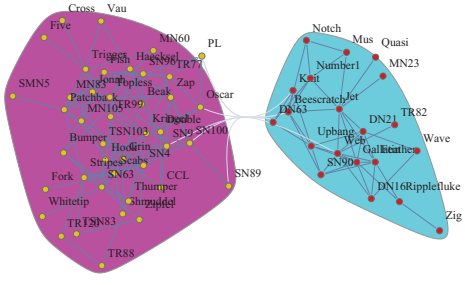
Fig. 5. Zechary Karate Club data.



(a) Network

Nodes	62
Edges	159
Nodes in largest Weakly Connected Component	62
Edges in largest Weakly Connected Component	159
Nodes in largest Strongly Connected Component	62
Edges in largest Strongly Connected Component	159
Diameter	8
Avg. clustering coefficient	0.258958

(b) Statistics



(c) Known Community Plots

Fig. 6. Dolphin Social Graph data.

from the granule center increases. Although these are usual assumptions in social network analysis, sometimes these may not be true depending on the data set. However, such characteristics may be accommodated in the said framework of knowledge representation just by changing the membership functions appropriately. For example, if the network contains multiple relationships, then one may assign memberships using a vector or multi-variable based distance function instead of a simple hop distance.

In designing our framework, we assumed the same role for all the actors in a network. This means, the model is valid for any social network as long as the roles of all the actors in the network remain the same. However, if a network contains different roles for its different

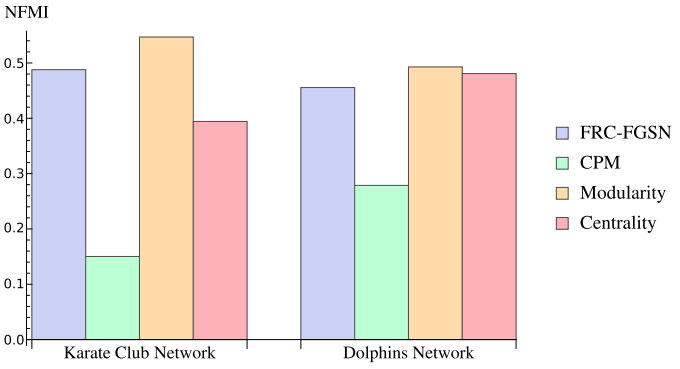


Fig. 7. Bar chart showing the comparative values of NFMI for different algorithms of Karate Club Network and Dolphin Social Network.

actors, then a modification may be required, in the proposed model, to accommodate such role-players in community structures.

Acknowledgments

The authors acknowledge the Department of Science and Technology, Government of India for funding the Center for Soft Computing Research at Indian Statistical Institute. S.K. Pal acknowledges the J.C. Bose National Fellowship and INAE Chair Professorship.

References

[1] J.L. Moreno, H.H. Jennings, in: Who Shall Survive? A New Approach to the Problem of Human Interrelations, Nervous and Mental Disease Monograph Series, Nervous and Mental Disease Publishing Co., New York, 1934.
[2] S. Fortunato, Community detection in graphs, Phys. Rep. 486 (3–5) (2010) 75–174.
[3] S.K. Pal, S. Kundu, C.A. Murthy, Centrality measures, upper bound, and influence maximization in large scale directed social networks, Fundam. Informatic. 130 (3) (2014) 317–342.
[4] F.D. Malliaros, M. Vazirgiannis, Clustering and community detection in directed networks: a survey, Phys. Rep. 533 (4) (2013) 95–142.
[5] A.L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (5439) (1999) 509–512.
[6] M. Faloutsos, P. Faloutsos, C. Faloutsos, On power-law relationships of the internet topology, in: Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication, SIGCOMM '99, ACM, New York, Cambridge, 1999, pp. 251–262.
[7] S. Chattopadhyay, C.A. Murthy, S.K. Pal, Fitting truncated geometric distributions in large scale real world networks, Theor. Comput. Sci. 551 (2014) 22–38.
[8] B. Krishnamurthy, J. Wang, On network-aware clustering of web clients, in: Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication, SIGCOMM '00, ACM, New York, Stockholm, 2000, pp. 97–110.
[9] P.K. Reddy, M. Kitsuregawa, P. Sreekanth, S.S. Rao, A graph based approach to extract a neighborhood customer community for collaborative filtering, in: Proceedings of the Second International Workshop on Databases in Networked Information Systems, DNIS '02, Springer-Verlag, London, 2002, pp. 188–200.
[10] M. Steenstrup, Cluster-based Networks, in: Ad Hoc Networking, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2001, pp. 75–138.
[11] S.A. Rice, The identification of blocs in small political bodies, Am. Politic. Sci. Rev. 21 (3) (1927) 619–627.
[12] R.S. Weiss, E. Jacobson, A method for the analysis of the structure of complex organizations, Am. Sociol. Associat. 20 (6) (1955) 661–668.
[13] M. Girvan, M.E.J. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. U.S.A. 99 (12) (2002) 7821–7826.
[14] M. Newman, M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E 69 (2) (2004) 1–15.
[15] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, second ed., Springer Series in Statistics, Springer, New York, 2009.
[16] U. Raghavan, R. Albert, S. Kumara, Near linear time algorithm to detect community structures in large-scale networks, Phys. Rev. E 76 (3) (2007) 36106.
[17] M. Newman, Fast algorithm for detecting community structure in networks, Phys. Rev. E 69 (6) (2004) 066133.
[18] A. Lancichinetti, S. Fortunato, Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities, Phys. Rev. E 80 (1) (2009) 016118.
[19] W. Zachary, An information flow model for conflict and fission in small groups, J. Anthropol. Res. 33 (4) (1977) 452–473.
[20] D. Lusseau, K. Schneider, O.J. Boisseau, P. Haase, E. Slooten, S.M. Dawson, The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations, Behav. Ecol. Sociobiol. 54 (4) (2003) 396–405.
[21] L.A. Zadeh, Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic, Fuzzy Sets Syst. 90 (1997) 111–127.
[22] S.K. Pal, B. Uma Shankar, P. Mitra, Granular computing, rough entropy and object extraction, Patt. Recog. Lett. 26 (16) (2005) 2509–2517.
[23] A. Lancichinetti, S. Fortunato, J. Kertész, Detecting the overlapping and hierarchical community structure in complex networks, New J. Phys. 11 (3) (2009) 033015.
[24] P. Maji, S.K. Pal, Fuzzy-rough sets for information measures and selection of relevant genes from microarray data., IEEE Trans. Syst., Man, Cybern. Part B 40 (3) (2010) 741–752.
[25] A. Lancichinetti, S. Fortunato, F. Radicchi, Benchmark graphs for testing community detection algorithms, Phys. Rev. E. 80 (1) (2008) 016118.
[26] G. Palla, I. Derényi, I. Farkas, T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society, Nature 435 (7043) (2005) 814–818.