Granular Social Network: Model and Applications

Sankar K. Pal and Suman Kundu

Abstract Social networks are becoming an integral part of the modern society. Popular social network applications like Facebook, Twitter produces data in huge scale. These data shows all the characteristic of Big data. Accordingly, it leads to a deep change in the way social networks were being analyzed. The chapter describes a model of social network and its applications within the purview of information diffusion and community structure in network analysis. Here fuzzy granulation theory is used to model uncertainties in social networks. This provides a new knowledge representation scheme of relational data by taking care of the indiscernibility among the actors as well as the fuzziness in their relations. Various measures of network are defined on this new model. Within the context of this knowledge framework of social network, algorithms for target set selection and community detection are developed. Here the target sets are determined using the new measure granular degree, whereas it is granular embeddedness, together with granular degree, which is used for detecting various overlapping communities. The resulting community structures have a fuzzy-rough set theoretic description which allows a node to be a member of multiple communities with different memberships of association only if it falls in the (rough upper - rough lower) approximate region. A new index, called normalized fuzzy mutual information is introduced which can be used to quantify the similarity between two fuzzy partition matrices, and hence the quality of the communities detected. Comparative studies demonstrating the superiority of the model over graph theoretic model is shown through extensive experimental results.

1 Introduction

Social network is a collection of social ties among friends and acquaintances. After a child is born, (s)he gets immediately connected with the members of the family. Over the course of time (s)he develops connections with larger networks like village,

S.K. Pal · S. Kundu (✉)
Center for Soft Computing Research, Indian Statistical Institute,
Kolkata 700108, India
e-mail: suman@sumankundu.info

A.Y. Zomaya and S. Sakr (eds.), *Handbook of Big Data Technologies*,
DOI 10.1007/978-3-319-49340-4_18

school, and office. Due to the technological advancement, distance travel, global communication and digital interaction have been growing in numbers and in effect social networks are also growing steadily in complexity. This complex "connectedness" of modern society got the attention of different fields of studies.

The term "social network" was coined by the social scientists. The network was considered as a theoretical construct to study the relationships between individuals, groups, organizations or even the entire society. However, the recent boom in online services related to social networks, viz Facebook, Twitter, WhatsApp, LinkedIn, provides new research opportunities to the scholars of computer science, because the data available from these networks are dynamic, large, diverse and complex. That is, it shows all the characteristics of Big Data [69] such as Velocity, Volume, and Variety. Accordingly, recent algorithms [43, 53, 61, 85] are addressing the Big Data issues related to social networks.

Since its inception in early 20^{th} century, social networks are represented using graphs [58], and graph analysis has become crucial to understand the features of these networks [24]. Due to the recent revolution in computing (processing) power, one can now handle relatively larger real networks [67] potentially reaching millions of vertices. Accordingly, it leads to a deep change in the way social networks were being analyzed.

In contrast to random network, social networks shows fascinating patterns and properties [57]. The degree distribution follows power law [5, 21] or truncated geometric distribution [8]. Diameter of the network is very small compared to the size of the network, and the network possesses high concentration of edges in its certain parts forming groups. Such groups with high internal edge density within themselves and low between them characterizes the community structure (or clusters) of the network.

Two of the major research areas in Social Network Analysis (SNA) are (a) analysis of network values [16, 39, 96], and (b) community detection [9, 65]. The objective of the former is to analyze the relative importance of a node in the network. One of the major research application of this area is **target set selection**. In this problem, one seeks to find a set of influential nodes for which the information diffusion over the network is maximum. It is effectively used in viral marketing [81] through online social networks. In addition, this can be used for finding the top stories from a news network, spreading of social awareness or combat with deceptions spreading over social media. Other applications of network values include study on epidemic spreading, diffusion of innovations, homophily analysis and optimal price-setting in market.

Several attempts [16, 35, 37, 38, 40, 76, 81, 95] were made to solve the target set selection problem. However, these are very restrictive either in terms of performance or in execution time, specifically for large scale social networks. For example, greedy hill climbing algorithm of [37] provides approximation within a factor of $(1 - \frac{1}{e} - \epsilon)$ to the optimal solution. Here e is the base of natural logarithm, and ϵ depends on the accuracy of Monte-Carlo estimate of influence spread. But it takes days to compute the set of seeds, even for a moderate sized social network. In contrast to this, heuristic methods (e.g., [10, 11]) are fast but provides sub-optimal output as compared to the greedy method of [37].

Community detection, on the other hand, deals with the problem of identifying virtual groups in a network. A community is formed when a group of nodes are more densely connected with each other compared to rest of the network. In addition to the social implication study of such groups, the solution to this problem has broad application in different fields. For example, in world wide web it will help to optimize the Internet infrastructure [42], in a purchase network it can boost the sell by recommending the appropriate products [78], and in computer network it will help to optimize the routing table creation [84].

Scientists from several disciplines studied the community detection problem for a long time [28, 54, 62, 63, 77, 80, 89]. These involve mainly two strategies for finding different 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, for large-scale networks, it is possible that a node may belong to more than one community with different degrees of association.

Beside these, highly overlapping neighborhoods in real life big social networks enforce uncertainties in decision making. Although the graph modeling has been in use for social networks since its inception in 1934 [58], a better modeling to deal with these uncertainties is in need. The new modeling may lead to a deep change in the way social networks were being analyzed.

2 Preliminaries

2.1 Social Network Analysis

At a more precise level, a network is any collection of objects in which some pairs are connected by links [17]. Based on configuration, different forms of relationships or connections may be used to define links. Due to this flexible options, it is easy to find network in different domains. Graph based modeling is a typical way to represent social networks. Let us first explain some of the basic elements of graphs before providing a review on modeling social network.

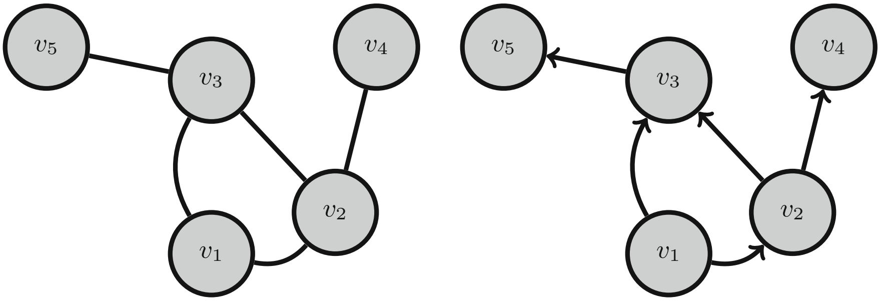
Graph, Nodes and Edges: Conceptually, a graph is formed by nodes (vertices) and edges (links) connecting the nodes. Formally, a graph is an ordered pair (V, E) where V is the set of nodes and E is the set of edges, formed by pairs of nodes.

Undirected and Directed Graphs: Edges can be symmetric such as in Fig. 1a, or asymmetric like in Fig. 1b. The former is referred as undirected graph or simply graph and the latter is called directed graph.

Graphs as Models of Networks: Graphs are useful in social network study as they serve as mathematical models of network structure. Let us now replace aforesaid toy example Fig. 1 with a real social network of Fig. 2. It is popularly known as Zachary karate club [92]. This network shows the friendship relations between 34 members of a US karate club in 1970s. People are represented by nodes and edges are constructed where two people shows friendship outside the context of club. Note that the actual placement of nodes is immaterial. All that matters is which node is connected with which others. Statistics about the network is shown in Table 1.

Paths and Cycles: A path is a sequence of nodes where each consecutive pair in the sequence is connected by an edge. For example, in the Zachary karate club we have a path from node 1 to 34 as 1, 14, 3, 34. A path can repeat nodes such as, 1, 4, 13, 1, 12. Cycle is a specific kinds of path which forms a ring like structure. For example, in Zachary karate club 11, 5, 7, 6, 11 is a cycle.

Connectivity: Whether we are dealing with small or large scale social networks, it is natural to check if every node can reach every other node via a path. We say a graph is connected if for every pair of nodes there exists a path between them. For any social network it may happen that two persons are not reachable via a valid path. This then leads to a disconnected network. For example, Fig. 3 shows a disconnected network of metabolic cellular network.



(a) Undirected Graph

(b) Directed Graph

Fig. 1 Example: graphs, nodes, edges

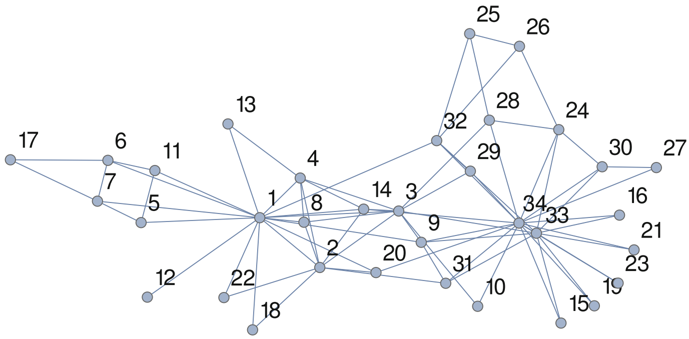


Fig. 2 Zachary karate club

Table 1 Statistics of Zachary karate club network

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

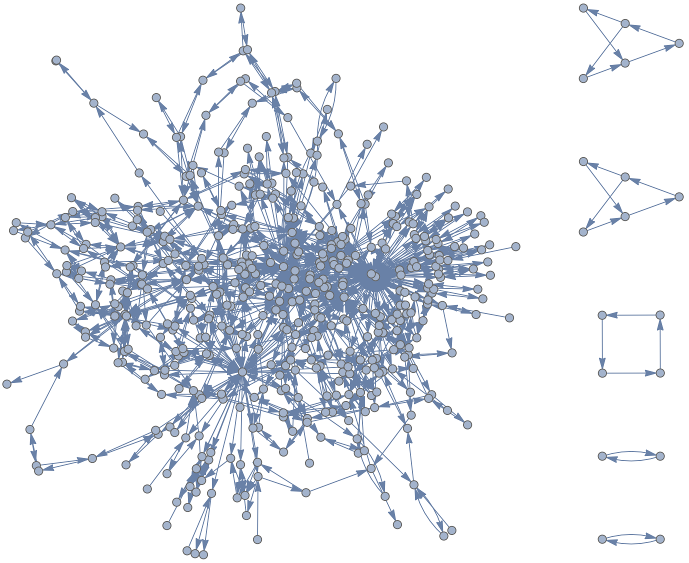
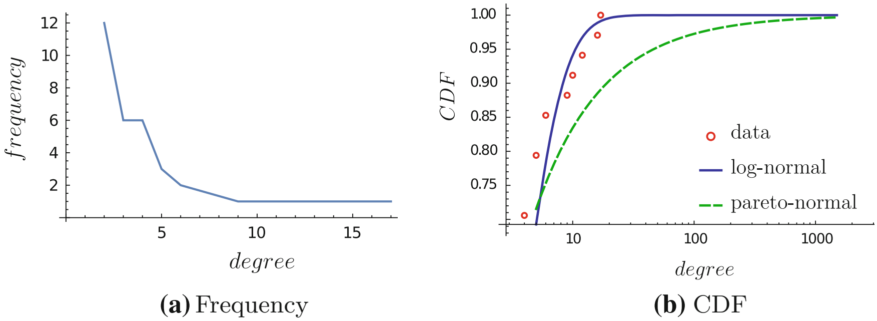


Fig. 3 Metabolic cellular network data for Oryza Sativa [33]

Components (Weakly Connected versus Strongly Connected) If a graph is not connected it breaks apart naturally. These separate subsets are called components. Each of the components when considered separately represents a connected graph. For example the disconnected network in Fig. 3 has 6 connected component.

For directed social network the notion of connectivity can be expressed in two different forms, namely, weakly connected component and strongly connected component. A weakly connected component is a subgraph of a directed graph such that for every pair of nodes u, v in the subgraph, there is an undirected path from u to v and a directed path from v to u. On the other hand, a strongly connected component is a subgraph of a directed graph such that for every pair of nodes u, v in the subgraph, there is a directed path from u to v and a directed path from v to u.



(a) Frequency **(b)** CDF

Fig. 4 Degree distribution of Zachary karate club network

Neighbors and Hop Distance: Two nodes u and v are said to be neighbors or adjacent when they are connected with an edge, i.e., (u, v) is a valid edge in the graph. If two nodes are not adjacent the distance along a path is usually measured by hop count. Hop count refers to the number of nodes one needs to pass from the source node to the destination node. That is a hop is one portion of the path from source to destination.

Shortest Path and Diameter: One may reach to a node u from another node v through different paths in the network. Shortest among them has significant value in the network study. A path p between nodes u and v is said to be the shortest if no other path between them in the network holds lesser length (in terms of hop distance) than p.

The diameter of a network is the length of the longest of the shortest paths in the network. In other words, among the all pairs shortest paths, the highest hop distance is the network diameter. For real world social networks, it is found that the diameters tend to be very small. For example, the diameter of the karate club network (Fig. 2) is 5. This phenomenon is called small world property of the social network.

Degree and Degree Distribution: Degree of a node is measured by the number of incident edges on it. It is denoted by $d(v)$. For directed graphs, a node has two different degrees, the in-degree, which is the number of incoming edges, and the out-degree, which is the number of outgoing edges.

Degree distribution refers to the frequency distribution of the degrees of a network. Degrees are usually plotted in x-axis and the frequencies are plotted in y-axis. Figure 4a shows the degree distribution of the karate club data. Similarly we can plot cumulative distribution function (CDF) as shown in Fig. 4b.

An observation can be made from the degree distribution of the karate club data that the number of nodes with higher degree is low as compared to the number of nodes with lower degree values. Similar long tail can be found in most of the real world networks. This is different from random graphs and due to this, social networks are referred as scale free network.

2.2 Fuzzy Sets

Traditional set theory deals with whether an element "belongs to" or "does not belong to" a set. Fuzzy set theory [93], on the other hand, concerns with the continuum degree of belonging, and offers a new way to observe and investigate the relation between sets and its members. It is defined as follows:

Let X be a classical set of objects, called the universe. A fuzzy set A in X is a set of ordered pairs $A = \{(x, \mu_A(x)) | x \in X\}$, where $\mu_A : X \rightarrow M$ is called the membership function of x in A which maps X to membership space M. Membership $\mu_A(x)$ indicates the degree of similarity (compatibility) of an object x to an imprecise concept, as characterized by the fuzzy set A. The domain of M is [0, 1]. If $M = \{0, 1\}$, i.e., the members are only assigned either 0 or 1 membership value, then A possesses the characteristics of a crisp or classical set.

The set of all elements having positive memberships in fuzzy set A constitutes its support set, i.e.,

$$Support(A) = \{x | \mu_A(x) > 0\}. \tag{1}$$

The cardinality of the fuzzy set A is defined as

$$|A| = \sum_{x \in X} \mu_A(x). \tag{2}$$

Union and intersection of two fuzzy sets A and B are also fuzzy sets and we denote them as $A \cup B$ and $A \cap B$ respectively. The membership functions characterizing the union and intersection of A and B are as follows:

$$\mu_{A \cup B}(x) = \max(\mu_A(x), \mu_B(x)), x \in X \tag{3}$$

$$\mu_{A \cap B}(x) = \min(\mu_A(x), \mu_B(x)), x \in X. \tag{4}$$

2.3 Rough Sets

Let X be a classical set of objects, in a universe of discourse U. Under situations when relations exist among elements of U, X might not be exactly definable in U as some elements of U that belong to the set X might be related to some elements of U that do not belong to set X.

When a relation, say R, exists among elements of U, limited discernibility draws elements of U together governed by the relation R resulting in the formulation of granules in U. Here, a set of elements in U that are indiscernible from or related to each other is referred to as a granule. Let us represent granules using Y and the family of all granules formed due to the relation R using U/R.

As mentioned earlier, the relation R among elements of U might result in an inexact definition of X. To tackle such cases, in rough set theory, X is approximately represented by two exactly definable set $\underline{R}X$ and $\overline{R}X$ in U given as

$$\underline{R}X = \bigcup \{Y \in U/R | Y \subseteq X\} \tag{5}$$

$$\overline{R}X = \bigcup \{Y \in U/R | Y \cap X \neq \} \tag{6}$$

In the above, the set $\underline{R}X$ is defined by the union of all granules that are subsets of the set X and the set $\overline{R}X$ is defined by the union of all granules that have non-empty intersection with the set X. The sets $\underline{R}X$ and $\overline{R}X$ are respectively called the lower approximation and upper approximation of X with the imprecise concept R.

Fuzzy set and rough set are reputed to handle uncertainities arising from overlapping concepts (or characters) and granularity in the domain of discourse respectively. While the former uses the notion of class membership of an element, the latter hinges on the concept of approximating from lower and upper side of a set defined over a granular domain.

2.4 Granular Computing

Granular computing (GrC) is a problem solving paradigm with the basic element, called granules. The construction of granules is a crucial process, as their sizes and shapes are responsible for the success of granular computing based models. Further, the inter and intra relationships among granules play an important role. A granule may be defined as the clump of elements that are drawn together, for example, by indiscernibility, similarity and functionality. Each of the granules according to its shape and size, and with a certain level of granularity may reflect a specific aspect of the problem. Granules with different granular levels may represent a system differently.

Granulation is the process of construction, representation and interpretation of granules. It involves the process of forming larger objects into smaller and smaller into larger based on the problem in hand. According to Zadeh [94], "granulation involves a decomposition of whole into parts. Conversely, organization involves an integration of parts into whole."

One of the realizations behind GrC is that - precision is sometimes expensive and not very meaningful in modeling and controlling complex systems. When a problem involves incomplete, uncertain and vague information, it may sometimes become difficult to differentiate the individual elements, and one may find it convenient to consider granules to represent a structure of patterns evolved by performing operations on the individual patterns [26]. Accordingly, GrC became an effective framework in designing efficient and intelligent information processing systems for various real life decision-making applications. The said framework can be modeled, for example, with the principles of fuzzy sets, rough sets, neural networks, power algebra, interval analysis [73]. For further details on the significance and various applications of GrC, one may refer to [7, 66, 68, 70, 72, 91].

3 Literature Review

3.1 Modeling Social Networks

As mentioned in Sect. 2.1, network structures with actors and their relationships are usually modeled as graphs. In sociology, this representation is sometime referred as sociogram. In a sociogram, actors are represented by vertices of a graph, and relations by edges. Graphs appear naturally here as it is useful to represent how things are either physically or logically linked together. Sociogram was developed by Moreno [58] to analyze the choices of preferences within a group. It was used to diagram the structure and patterns of group interactions.

Social network data, sometime represented in two-way matrices, is termed as sociomatrices [88]. The two dimensions of a sociomatrix are indexed by the senders (rows) and the receivers (column) of relationships. Usually the matrix has n rows and n columns, where n represents the number of actors in the network. Thus a basic sociomatrix is square. Sociomatrices were first used together with sociogram by Moreno [58] who showed how social relationship can be pictured through these.

The same network can also be represented using the relational form. Relational algebras (also called role algebras) are used to analyze the structure of social roles by emphasizing multiple relations rather than actors. Harrison White and his students [6, 90] pioneered this approach as an extension to block modeling. A block model is a representation of objects in groups based upon patterns that occur in the relations between these objects [3]. The structure of a block model is a matrix in which the $(i, j)^{th}$ entry denotes the number of directed edges from nodes in cluster i to nodes in cluster j. A block models can represent any pattern that arises in the relations between objects, such as bipartite relations, hierarchies, rings, bridges, and other unique aggregate connectivity patterns between groups of vertices.

Another approach to model social networks is based on statistical modeling. The idea of statistical modeling of network is to represent the main features of the social network by a few parameters and express the uncertainty of those estimates by standard error, p-value, posterior distribution etc. There are two ways for statistical modeling of network, viz. model-based inference and design-based inference. When a sample is drawn from a larger graph, design-based method can be used. Link-tracing technique [83] is one kind of design based method. Examples of this technique are snowball design and random walk design. On the other hand, in model-based inference, it is required to construct a probability model with the assumption that the observed data can be regarded as the outcome of a random draw from this model [25, 27]. Multiple linear regression models are an example.

Thus several models for describing social network exist starting from 1930s. Recently, the development on modeling social network problems using multi-agent theory and/or game theory has been observed. In their paper [41], Kleinberg et. al. modeled a network with n distinct agents who build link to one another based on a strategic game. The payoff to an agent arises as a difference of costs and benefits. Narayanam et. al. [60], on the other hand, mapped the information diffusion process of social network to the formation of coalitions in an appropriately defined cooperative game. In [34], authors modeled the user interactions of a network to explore the dynamic evolutionary process of knowledge sharing among users using the agent-based computational approach. But the focus of these researches is mostly problem centric.

Fuzzy set theory has also received attention on social network analysis in recent years. In their work, Nair and Sarasamma [59] analyzed multi-modal social networks using fuzzy graphs and referred it as fuzzy social network. Later in 2008, Davis and Carley [14] used a stochastic model to identify fuzzy overlapping groups in social networks. Here they modeled the fuzzy overlapping group detection using an optimization problem. Another area where fuzzy sets have been used by different scientists is positional analysis (finding similarities between actors in the network) of social networks [22]. Instead of a general framework, these recent developments of fuzzy set theoretic approach in social network are more focused on a particular type of the network or particular application of the network.

Beside these, an attempt was made to use the concept of granular computing to model relational database for association discovery [32]. The technique is a specialized version of the general relational data mining framework which efficiently provides the search space for association discovery. Also, there were several research investigations focused on a problem oriented modeling of social network using different soft computing tools. For example, Chen and Li [9] proposed evolutionary computing based algorithm to detect community structures in complex networks. Genetic algorithm based diffusion model for information cascade in a social network is used in [46, 52]. For target set selection problem, Wang et al. [86] proposed a set-based coding genetic algorithm. However, none of these techniques provides any general framework which can serve as a generic platform, similar to sociogram or sociomatrices, to analyze social network data in view of different problems in the field.

Algorithm 1: Greedy Hill Climbing Algorithm

```
input          : A Social Network $G(V, E)$ and $k$
output         : Set $S \in 2^V$ having cardinality $k$
initialization: $S := \emptyset$
while $|S| \neq k$ do
    $v^* \leftarrow \arg\max_{v \in V \setminus S} \hat{\sigma}(S \cup \{v\})$ ; /* $\hat{\sigma}(.)$ returns the estimated influence */
    $S \leftarrow S \cup \{v^*\}$;
```

3.2 Target Set Selection

In the area of information diffusion, finding a target set is to find the influential nodes mainly in terms of the total influence in the network. The natural solution to the problem will be to select those persons having higher numbers of neighbors. That is, select the nodes based on their degree centrality scores. Domingos and Richardson were the first to study the problem [16, 81] in the algorithmic aspect and proposed probabilistic methods to solve it. Later, Kempe et al. formulated it as a discrete optimization problem [37] and showed that the problem is NP hard. They proposed a greedy hill climbing algorithm shown in Algorithm 1. In each iteration of the algorithm, marginal contribution of every non seed node (i.e., nodes in $V \setminus S$) to the information diffusion is separately estimated and the highest contributor is selected as the next seed. Thus the algorithm maximizes the influence contribution during seed selection. Hence it is able to find higher quality seeds. However, for the same reason it leads to high computational time, specially for large scale networks. The main drawback of the algorithm comes from the marginal contribution estimation. There is no deterministic methods available till date to get the marginal contribution of a node. In their paper, Kempe et al. [37] uses Monte Carlo simulation for the estimation os such contribution. As the process of information diffusion is highly stochastic, the simulation needs to be performed for a sufficiently large number of times to obtain more accurate results. It may take days to identify top 50 seeds even on a graph of moderate size of 30 K nodes [12]. To overcome this drawback, several algorithms were proposed in last few years [11, 18, 30, 49]. Notably, in [49] Leskovec et al. presented a *cost-effective lazy forward* (CELF) method which exploits the sub-modularity property of the influence function. For any given set function $\sigma(.)$, sub-modularity property confirms that the effect of v to a subset is always higher than that of the super set. That is, $\sigma(S \cup \{v\}) > \sigma(T \cup \{v\})$ if $S \subset T$. Authors argued in [49] that most of the realistic outbreak detection objectives are *sub-modular*. Their experiments with blog network and water network show that CELF runs 700 times faster than the greedy algorithm of [37]. However, CELF method still takes hours to generate 50 seeds [11]. Improvement in execution time was also sought by considering the properties of the underlying diffusion model. One of such popular diffusion models is Independent cascade model. In this model of information diffusion, information propagates in discreet time steps. In each time t, one node with the information tries to influence one of its neighbors who does not have the information already. Success depends on a probability called propagation probability. Irrespective of the success, the same node will never get a chance to influence the same neighbor again. In [11], authors provide two new greedy algorithms designed on independent cascade model of information diffusion. One of them, NewGreedyIC, uses a random removal of edges instead of Monte Carlo simulation to estimate a node's marginal contribution. The random removal uses the propagation probability to identify the edges to be removed. This process leads to an improvement in execution time. Further they integrated the idea of the CELF inside the NewGreedyIC and proposed improved MixGreedyIC. Goyal et al. [30] suggested an improved version of CELF as CELF++ and showed empirically that the algorithm is faster than CELF with insignificant amount of additional memory usage. In CELF++, authors maintained a heap with intermediary results of the Monte Carlo simulation, which reduces the execution time of the subsequent iterations. A greedy sketch-based influence maximization (SKIM) was described in [13] very recently and it is reported that it may be scaled to large social network data.

In contrary, several heuristic algorithms [10–13] were proposed which improve the performance compared to the centrality measures while the execution time remains lower than that of the greedy. One such algorithm is degree discount heuristic of [11], which runs with the following principle. If a node u is already considered as a seed then in later iterations a node v's degree is calculated after discounting the edge $e(u, v)$. This algorithm works very well for undirected social networks. In 2012, Wang et al. [87], reported a heuristic method named prefix excluding maximum influence path (PMIA) where the propagation probability of a path is calculated and used to identify a node's contribution in the diffusion. These heuristic approaches used underlying diffusion principles to improve the performance. Some of the heuristic algorithms, on the other hand, are designed to perform well on specific social networks. For example, Chen et al. [12] proposed a liner time algorithm for directed acyclic graphs, and Gomez- Rodriguez and Schölkopf [29] proposed probability based methods to identify influential nodes for continuous time diffusion networks. Similarly, Aslay et al. [4] described a target set selection algorithm for topic-aware influence maximization queries and Li et al. [51] reported a location-aware target set selection method using spacial-based indexes.

3.3 Community Detection

Community detection is to identify virtual groups of a network. The main challenge is to identify the groups and possibly their hierarchical organization by only using the network topology. One of the first studies on community identification was carried out by Rice [80]. In the work, clusters were identified in a small political body based on their voting patterns. Later in 1955, Jacobson [89] studied community structure within a government agency [89]. They have separated work-groups by removing those people who work with different groups. This idea of removing edges is the basis of several algorithms in recent times. One such algorithm, presented by Girvan and Newman [28], aims at the identification of the edges lying between two communities for possible removal. These edges were identified based on their centrality values. The concept is considered as the start of modern era in community detection. In [63], Newman [63] proposed the modularity measure to quantify the quality of the identified community structure. The modularity is, up to a multiplicative constant, the number of edges falling within groups minus the expected number in an equivalent network with edges placed at random [64]. Modularity value can be either positive or negative. Positive value of modularity indicates the presence of community structures. So, one may partition the network with the aim to maximizing the modularity value of the community structure. This idea of optimizing modularity using some optimization technique is used to identify the community structure by Newman [62]. On the contrary, Raghavan et al. [77] described a near liner localized community detection algorithm based on label propagation which does not optimize any similar measures of community strength. In this method, initially each node is assigned with a unique label. At every iteration of the algorithm, each node adopts the label which is used by maximum number of its neighbors. Ties are broken randomly. At the end of the algorithm, nodes with the same label are grouped together to form a community. Density based graph partitioning algorithm is also available in the literature. Example of one such algorithm is by Falkowski et al. [20]. More traditional methods such as hierarchical [31] and partition based clustering, where vertices are jointed into groups as per their mutual similarities, are also used for identifying communities in a social network.

All these algorithms discussed above create a crisp partition in the network. That is, a node belongs to a single community only. However in a real life a person may belong to multiple groups, i.e., the existence of overlapping community structures. In [36, 79], authors showed that overlapping is indeed a significant feature of many real world social networks. One of the most popular overlapping community detection algorithms, namely, clique percolation method (CPM) of [71], detects overlapping communities by searching of adjacent cliques. The algorithm first searches for all the cliques of size k and constructs another graph by considering a k-clique as a node. A link is added when two cliques share $k - 1$ edges. Each connected component on this new graph is considered to be a community in the network, and k-cliques belonging to a component are considered to be in the same community. Overlap is possible because a node can be a member of multiple cliques. A version of the same algorithm for weighted network was proposed by Farkas et al. [23]. Here k-cliques with weight greater than a threshold are considered for the community. Another approach to get overlapping community structure is to partition links instead of nodes. Ahn et al. [2] used hierarchical clustering to partition edges of the network. In this algorithm, each edge belongs to a unique cluster but nodes may naturally belong to different clusters. Evans and Lambiotte [19], on the other hand, constructed a new weighted line graph by considering links of the original graph as a node and then partition this new graph using disjoint community detection algorithm. Although the link partitioning for overlapping detection seems conceptually to be natural, there is no guarantee that it provides higher quality than the node based detection does [24]. Readers may refer to [24, 50] for review on different Community detection algorithms.

4 Fuzzy Granular Social Networks (FGSN)

Social network is nothing but a collection of relations between social actors and their interactions. Social actors often form closely operative groups among themselves, which are often indistinguishable. A granule is a clump of objects (points) in the universe of discourse, drawn together, for example, by indistinguishability, similarity, proximity or functionality [94]. So, the characteristic of indistinguishability among closely operative groups of the social actors may be modeled using the concepts of granules for further processing.

Further, the basic concepts of conceptual similarities between nodes, cluster of nodes, relations between nodes and their interactions etc. do not lend themselves to precise definition, i.e., they have ill-defined boundaries. So, it is appropriate and natural if a social network is viewed in terms of a collection of *fuzzy granules*. Based on these notions, a new unified framework to model social networks effectively and efficiently in the framework of granular computing is developed [44, 45]. In this model a granule is constructed around a node with fuzzy boundary. The membership function for computing the degree of belonging of a node to the said granule is determined depending upon the problem in hand. Within this framework, some of the popularly known network measures are redefined [44].

4.1 The Model

Global phenomenon of a social network always ensembles the local behaviors of individuals as well as their closely related neighborhoods. While the concept of neighborhoods in the network can be modeled in terms of granules, the vagueness in term "closeness" can be quantified using fuzzy set theory. In this section, we provide the description of the model *fuzzy granular social network* (FGSN).

Knowledge Representation: Let us consider the graph $G(V, E)$ represents a social network, where V is the set of all nodes (or vertices) and E represents the relationships (or edges). If I is the unit interval [0, 1], a fuzzy granular neighborhood defined over V is a function $\phi : V \rightarrow A(V)$, which assigns every node $v \in V$ to a fuzzy set $A \in I^V$. When $\phi(v)$ is non empty, we call it the fuzzy neighborhood of the node v, i.e., $\phi(v)$ is the granule defined around the node v. Due to the complex nature of social networks a node can be a member of different such neighborhood sets reflecting its different degrees of association. Let family of fuzzy sets associated with the node $v \in V$ be $\Phi(v)$. $\Phi(v)$ represents the neighborhood sets of node v. A fuzzy granular social network is represented by a triple:

$$\mathcal{S} = (\mathcal{C}, \mathcal{V}, \mathcal{G}) \text{ where} \tag{7}$$

- $\mathcal{V}$ is a finite set of nodes of the network
- $\mathcal{C} \subseteq \mathcal{V}$ is a finite set of granule representatives
- $\mathcal{G}$ is the finite set of all granules,

i.e.,$\mathcal{G} = \{\bigcup \Phi(c) | c \in \mathcal{C}\}$

A granule $g \in \mathcal{G}$ around a representative node ($c \in \mathcal{C}$) is constructed by assigning fuzzy membership values to its neighborhood nodes. Due to the overlapping nature of the neighborhoods, a node may belong to more than one granule. Their association with different granules may have different degrees as well. However, in case of directed social network, two different granules may be constructed around one single node. One for inbound relations and other for the outbound relationships [44].

4.2 Network Measures of FGSN

A social network is analyzed based on social measures defined over its graph representation. Similarly, several equivalent granular measures available for FGSN are provided in this section.

Let us first see the construction of FGSN of our example network shown in Fig. 2. Our objective here is to model the graph representation $G(V, E)$ by a *fuzzy granular social network* representation $\mathcal{S}(\mathcal{C}, \mathcal{V}, \mathcal{G})$. So, we need to define three sets $\mathcal{C}$, $\mathcal{V}$ and $\mathcal{G}$ from the network $G(V, E)$.

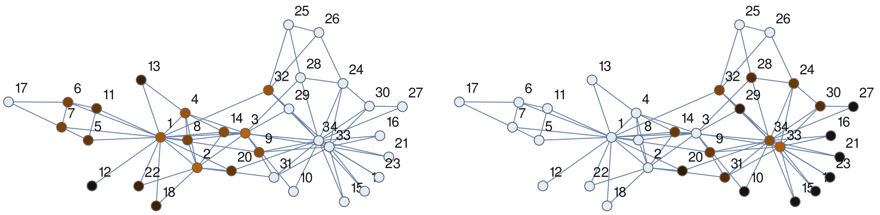
We consider preserving the maximum information of the network inside the FGSN. So, we constructed granules around every nodes in the network. Following is the definition of $\mathcal{S}(\mathcal{C}, \mathcal{V}, \mathcal{G})$ for Zachary karate club data.

- $\mathcal{V} = \{v | \forall v \in V\}$
- $\mathcal{C} = \{c | \forall c \in \mathcal{V}\}$
- $\mathcal{G} = \{A_c | \forall c \in \mathcal{C}, A_c \equiv \sum_{v \in \mathcal{V}} \tilde{\mu_c}(v)/v\}$.

Normalized membership value $\tilde{\mu_c}(v)$ is the degree of belonging of node v in the granule (A_c) around node c. $\tilde{\mu_c}(v)$ is calculated based on the Eq. 8 with minimum *hop* distance as the distance metric and $r = D$, the network diameter.

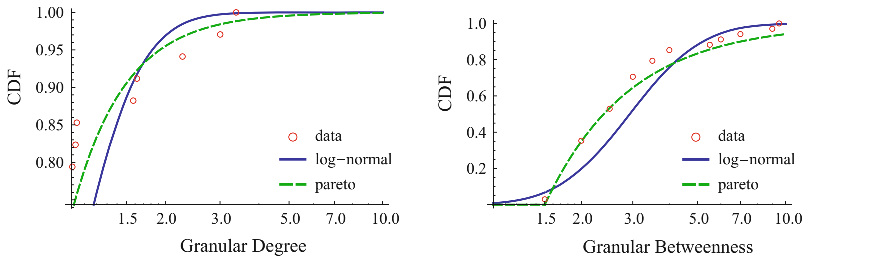
$$\tilde{\mu_c}(v) = \frac{\mu_c(v)}{\sum_{i \in \mathcal{C}} \mu_i(v)} \text{ such that } \sum_{i \in \mathcal{C}} \tilde{\mu_i}(v) = 1 \tag{8}$$

where,



(a) Granule of 1 **(b)** Granule of 34

Fig. 5 Color coded granules of Zachary karate club



(a) Granular degree distribution **(b)** Granular betweenness distribution

Fig. 6 FGSN features of Zachary karate club

$$\mu_c(v) = \begin{cases} 0 & \text{for } d(c, v) > r \\ \dfrac{1}{1 + d(c, v)} & \text{otherwise} \end{cases} \tag{9}$$

where $d(c, v)$ is the distance between node v and the center c.

Two such granules around nodes 1 and 34 are shown in Fig. 5. Here darker shades of brown represent higher values of membership. As we used normalized membership values, the nodes in less overlapping region may turn to have higher membership than the center nodes of the granules. This indicates that those nodes belong only to a fewer number of granules as compared to the centers. This is intuitively appealing as the former ones have higher possibilities of 'definitely belonging' to a granule than the latter ones

Granular Degree of a Node: *Granular degree* of a node in FGSN is equivalent to the degree measure of a node in graph representation. Granular degree of a node c is the cardinality of the granule constructed around the node c [44]. Here each granule is represented by a fuzzy set, so we use Eq. 2 to compute the granular degree of a node c as

$$\mathcal{D}(c) = |A_c| = \sum_{v \in \mathcal{V}} \tilde{\mu_c}(v) \tag{10}$$

In the karate club example (Fig. 2), node 34 has a granular degree of 3.38026 and node 1 has a granular degree of 3.0044. Figure 6a shows the distribution of granular degree of karate club data.

Granular Betweenness of a Node: *Granular betweenness* of a representative node c in FGSN is quantified by the sum of membership values that c possesses for all granules in the system [44]. Using the normalized membership values (Eq. 8), granular betweenness of $c \in \mathcal{C}$ is calculated as follows.

$$\mathcal{B}(c) = \frac{1}{\max\limits_{i \in \mathcal{C}} (\tilde{\mu_i}(c))} \tag{11}$$

$\mathcal{B}(c)$ takes values in $[0, |\mathcal{C}|]$. In our example karate club network, granular betweenness of node 1 and node 34 is 9 and 9.5, respectively. The distribution of granular betweenness of karate club data is shown in Fig. 6b.

Granular Embeddedness of a Pair of Nodes: *Granular embeddedness* for any pair of nodes defines how much a granule centered at one node is embedded inside that of the other [44]. It may be measured by the cardinality of the intersection of granules centered by the pair of points. Using Eqs. 4 and 2, granular embeddedness of a pair of nodes a and b is defined as

$$\mathcal{E}(a, b) = |A_a \cap A_b| = \sum_{v \in \mathcal{V}} \min(\tilde{\mu}_a(v), \tilde{\mu}_b(v)) \tag{12}$$

where A_a and A_b are the fuzzy sets representing the granules having the center nodes a and b, respectively.

In the example of karate club, the embeddedness of 1 and 34 is found to be 0.610714 when $r = 2$, and 0.959073 when $r = D(= 5)$, the diameter of the network.

4.3 Uncertainties in FGSN

Uncertainties in a social network arises due to the presence of vaguely defined *closeness* between nodes. Each relationship has a degree of togetherness. The presence of a relational link in a network does not imply that both the nodes are 100% committed towards each other. Similarly, the absence of a link does not necessarily mean they are not following each other. Let us now define two measures of uncertainties in FGSN in terms of fuzziness, as follows:

Energy Measure of a Granule in FGSN: Let us consider a monotonically increasing mapping $e : [0, 1] \rightarrow [0, 1]$ with the boundary conditions $e(0) = 0$ and $e(1) = 1$. An energy measure of a granule $A_c \in \mathcal{G}$, denoted by $\mathbb{E}(A_c)$, is a function of its characterizing membership values, represented as

$$\mathbb{E}(A_c) = \sum_{x \in \mathcal{V}} e[\tilde{\mu}_c(x)] \tag{13}$$

This measure quantifies the energy associated with the granule A_c. The energy increases as the membership values of its supporting nodes increase. The energy measure of A_c reduces to its cardinality if we use the identity mapping $e(x) = x \ \forall x \in \mathcal{V}$, i.e.,

$$\mathbb{E}(A_c) = \sum_{x \in \mathcal{V}} \tilde{\mu}_c(x) = |A_c| \tag{14}$$

One can also think of a different functional for e other than the identity mapping, for example, $e(x) = x^a, a > 0$ or $e(x) = sin(\frac{\pi}{2}x)$.

Entropy Measure of FGSN: Given a FGSN $\mathcal{S}(\mathcal{C}, \mathcal{V}, \mathcal{G})$, each granule $A_c \in \mathcal{G}$ represents a fuzzy equivalence class under the attribute set $\mathcal{C}$. If we have n objects in the universe $\mathcal{V}$ then the fuzzy relative frequency [56] of a granule will be

$$\rho(A_c) = \frac{|A_c|}{n} \tag{15}$$

where $|A_c|$ is the cardinality of the granule A_c. Based on this relative frequency of granules, one can find the information gain of the FGSN through its entropy, using Shannon's logarithmic function, as

$$H(\mathcal{S}) = -\sum_{A_c \in \mathcal{G}} \rho(A_c) log_\beta(\rho(A_c)) \tag{16}$$

where β represents the base of logarithm. Applying Eq. 15 into Eq. 16 we get

$$H(\mathcal{S}) = -\frac{1}{n} \sum_{A_c \in \mathcal{G}} |A_c| log_\beta(\frac{|A_c|}{n}). \tag{17}$$

The value of $H(\mathcal{S})$ can vary in $[0, log_\beta(|\mathcal{C}|)]$. $H(\mathcal{S}) = 0$ means the FGSN is least uncertain, while its value equal to $log_\beta(|\mathcal{C}|)$ signifies the highest uncertainty. Readers may refer [82] for generalized entropy measures in the granular space.

4.4 *Granular Degree Heuristic for Target Set Selection in FGSN*

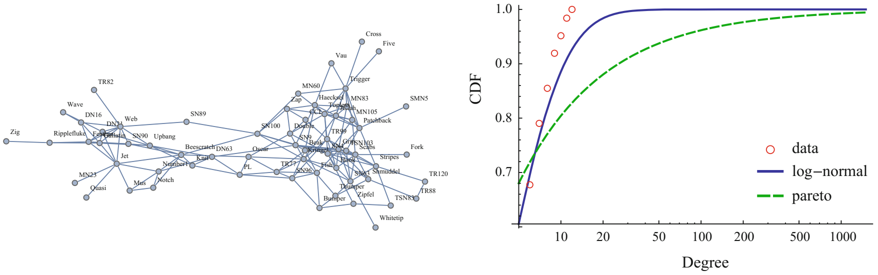
The section report the experimental results demonstrating the applicability of fuzzy granular social network for target set selection problem.

Problem Statement: Let us consider an influence function $\sigma : 2^{\mathcal{V}} \rightarrow \mathbb{N}$, defined for a social network $\mathcal{S}(\mathcal{C}, \mathcal{V}, \mathcal{G})$, such that given a set of initial active nodes $K \in 2^{\mathcal{V}}$, $\sigma(K)$ returns the expected number of active nodes at the end of information cascade. The problem of target set selection is to find the k number of influential nodes for which influence in $\mathcal{S}$ is maximum. So, this is a maximization problem defined as

$$\begin{aligned} \max_{K} \quad & \sigma(K) \\ \text{subject to} \quad & |K| = k, k > 0. \end{aligned}$$

Data Sets: In the experiments, we used three data sets, namely Zachary karate club [92], Dolphin social network [55], and Political blog network [1]. We already described the details of Zachary karate club in Sect. 2.1. Properties of Dolphin social graph and Political blog network are shown in Figs. 7 and 8 respectively.

Results: We first selected the top-k nodes (that is, the centers of the granules) from a given FGSN, in descending order of granular degree value. We refer this algorithm



(a) Network **(b)** Degree distribution

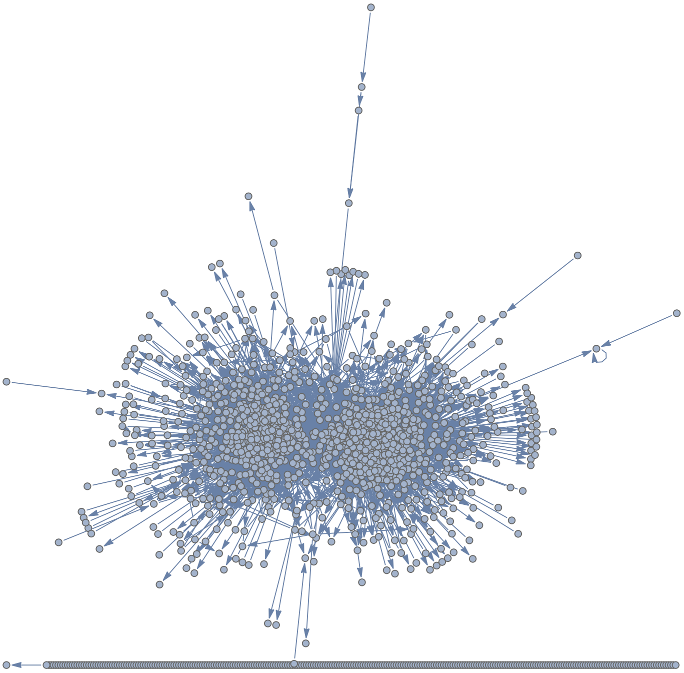
Nodes	62
Edges	159
NodesinlargestWeaklyConnectedComponent(WCC)	62
Edges inlargestWCC	159
NodesinlargestStronglyConnectedComponent(SCC)	62
Edges inlargestSCC	159
Diameter	8
Avg. clusteringcoefficient	0.258958

(c) Statistics

Fig. 7 Dolphin social graph

as Granular degree heuristic. Then we pass these top k nodes, as the set of seeds, in the Monte Carlo simulation of information diffusion (independent cascade model [37]). The output of the simulation process represents the number of total nodes influenced due to the said set of input seeds. We have varied the value of k from 1 to 15. These results are reported graphically in Fig. 9. Here X-axis shows the value of k and the Y-axis presents the total number of nodes influenced. As the Monte Carlo process is a stochastic process, we executed each experiment for 10000 trials and reported here the average values. It is clear from the figure that, for Zachary karate club and Dolphin social networks, results obtained with the proposed granular degree heuristic on FGSN outperform those obtained by other graph theoretic algorithms (High degree heuristic, Random and Diffusion degree heuristic [67]) for most values of k. This signifies that the set of seeds selected using the FGSN based method is able to determine the superior top k influential nodes. For Political blogs, the performance is at par with the High Degree Heuristic, superior to random and inferior to Diffusion Degree Heuristics.

Execution time (in seconds) of different algorithms for 1000 runs is shown in Table 2. As expected, the random selection method needs least time for all the data sets. Diffusion degree heuristics, on the other hand, takes longest time for all the cases. The proposed Granular degree heuristic requires much lower execution time as compared to diffusion degree for all the data sets. For Zachary karate club and Dolphin social graph, it is almost as fast as the high degree heuristics. For Political blog network, however, the proposed algorithm takes longer time compared to high degree heuristics. Further the algorithm is seen to perform best for $r = 2$. For other values e.g., $r = 1, 3, 4, 5$, the performance deteriorates [44].



(a) Network

Nodes	1490
Edges	16718
Nodes in largest Weakly Connected Component (WCC)	1222
Edges in largest WCC	16717
Nodes in largest Strongly Connected Component (SCC)	1
Edges in largest SCC	0
Diameter	10
Avg. clustering coefficient	0

(b) Statistics

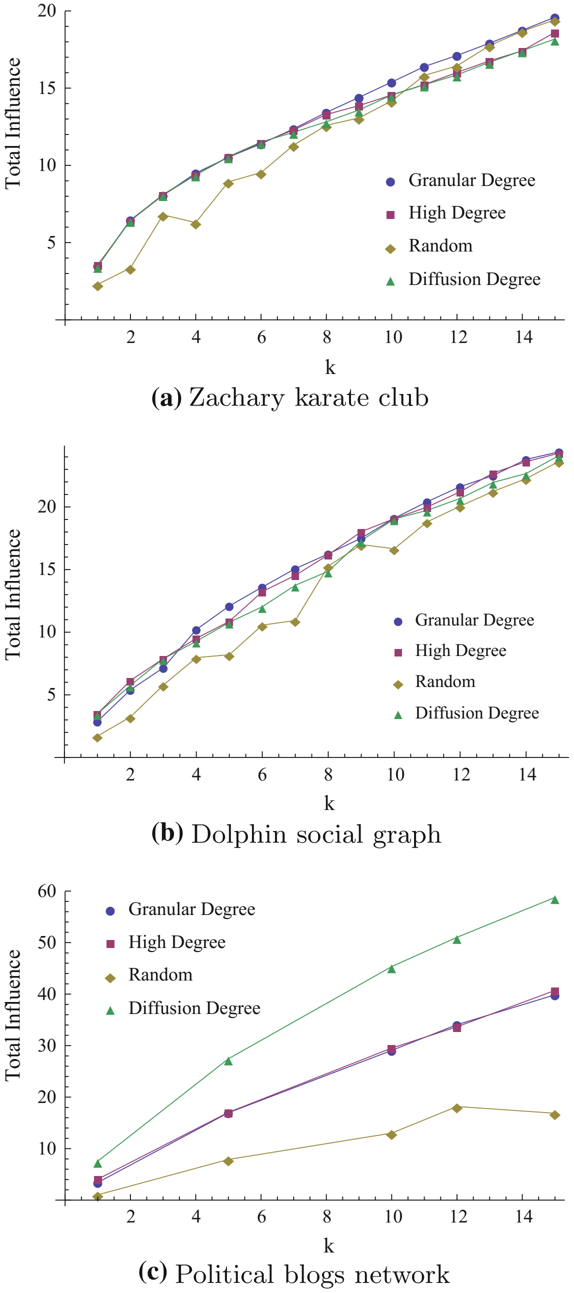
Fig. 8 Political blogs network

The computation complexity of the granular degree heuristic is $O(|V| + |E| + |\mathcal{C}| + kn)$ as reported in [44]. Here k is the number of desire seeds and n is the number of granules having granular degree greater than 1.

4.5 Fuzzy-Rough Community (FRC) Detection

A new community detection algorithm within the new knowledge representation scheme of FGSN is described in this section. Communities detected here show fuzzy-

Fig. 9 Variation of total influence with k for different algorithms ($r = 2$)



(a) Zachary karate club

(b) Dolphin social graph

(c) Political blogs network

Table 2 Execution time (in sec) of different algorithms for 1000 runs

Algorithms	Data Sets		
	Zachary karate club	Dolphin social graph	Political blogs network
Granular degree heuristics	0.311	0.52	27.26
High degree heuristics	0.3	0.48	3.7
Random selection	0.2	0.2	0.5
Diffusion degree heuristics	12.19	16.532	9.29×10^4

rough characteristics [45]. Nodes surely belong to a community constitute its lower bound (i.e., core region) in the notion of rough set theory while the others possibly belonging to the community are identified as members of "upper - lower" bound or boundary region. The nodes in the core region of the community are assigned with "unity" (full) membership to that community and "zero" (no) for the remaining community. The nodes in boundary region belong to multiple communities with different memberships of association. We assign fuzzy membership to these nodes based on their connectivity with different core regions, thereby resulting in unequal membership unlike the previous available methods.

The Algorithm: 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 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* (Eq. 10) exceeds a certain threshold (θ) indicating dense region. These granules are referred as θ-Core.

A community may have multiple such θ-cores. The algorithm needs to identify the set of those close by θ-cores. So, the goal is to search for θ-cores which belong to same community. These are called 'community reachable cores' [45]. To understand community reachability, we need to understand how the neighborhood of a granule is defined. Neighborhood of a granule A_c is the set of all granules whose centers lie in the support set of A_c [45], i.e.,

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

where $Support(A_c) = \{v | \tilde{\mu_c}(v, r) > 0\}$ and 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. There are three notions of community reachability. Two θ-cores are said to be (1) directly community reachable when one of them is in the neighborhood of the other, (2) indirectly community reachable, when one is reachable via a chain of directly community reachable θ-cores to the other and (3) r-connected community reachable when both of them are indirectly community reachable to a third θ-core r.

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, a normalized granular embeddedness measure [45] is introduced as

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

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

On the basis of community reachable θ-cores one may define community as follows.

Definition 1 (*Community*) Given a social network $\mathcal{S} = (\mathcal{C}, \mathcal{V}, \mathcal{G})$, and θ 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 reachable cores
- $\forall A_p \in \mathbb{C}, \mathcal{E}(A_p, \bigcup_{A_q \in \mathbb{C} \setminus A_p} A_q) > \frac{1}{\epsilon}$

θ and ϵ are referred as density and coupling co-efficient of the community respectively [45]. 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 2).

Definition 2 (*Fuzzy-Rough Community*) Let the n communities found for a social network be $\mathbb{C}_1, \mathbb{C}_2, ..., \mathbb{C}_n$, and the upper and lower approximation of the i^{th} community be $\overline{\mathbb{C}_i\theta}$ and $\underline{\mathbb{C}_i\theta}$ respectively. Then

$$\begin{aligned} \underline{\mathbb{C}_i\theta} &= \{x | x \in Support(A_p) \wedge x \notin Support(A_q); \\ &\qquad \forall A_p \in \mathbb{C}_i \text{ and } A_q \in \mathbb{C}_j; i \neq j\} \\ \overline{\mathbb{C}_i\theta} &= \{x | x \in Support(A_p); A_p \in \mathbb{C}_i\} \end{aligned} \tag{18}$$

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

$$\delta^{\theta}_{\mathbb{C}_i}(x, r) = \begin{cases} 1 & \text{if } x \in \underline{\mathbb{C}_i\theta} \\ \sum\limits_{c \in \underline{\mathbb{C}_i\theta}} \tilde{\mu_c}(x, r) & \text{if } x \in \overline{\mathbb{C}_i\theta} \setminus \underline{\mathbb{C}_i\theta} \\ 0 & \text{Otherwise} \end{cases} \tag{19}$$

where $\tilde{\mu_c}(x, r)$ is defined in Eq. 8.

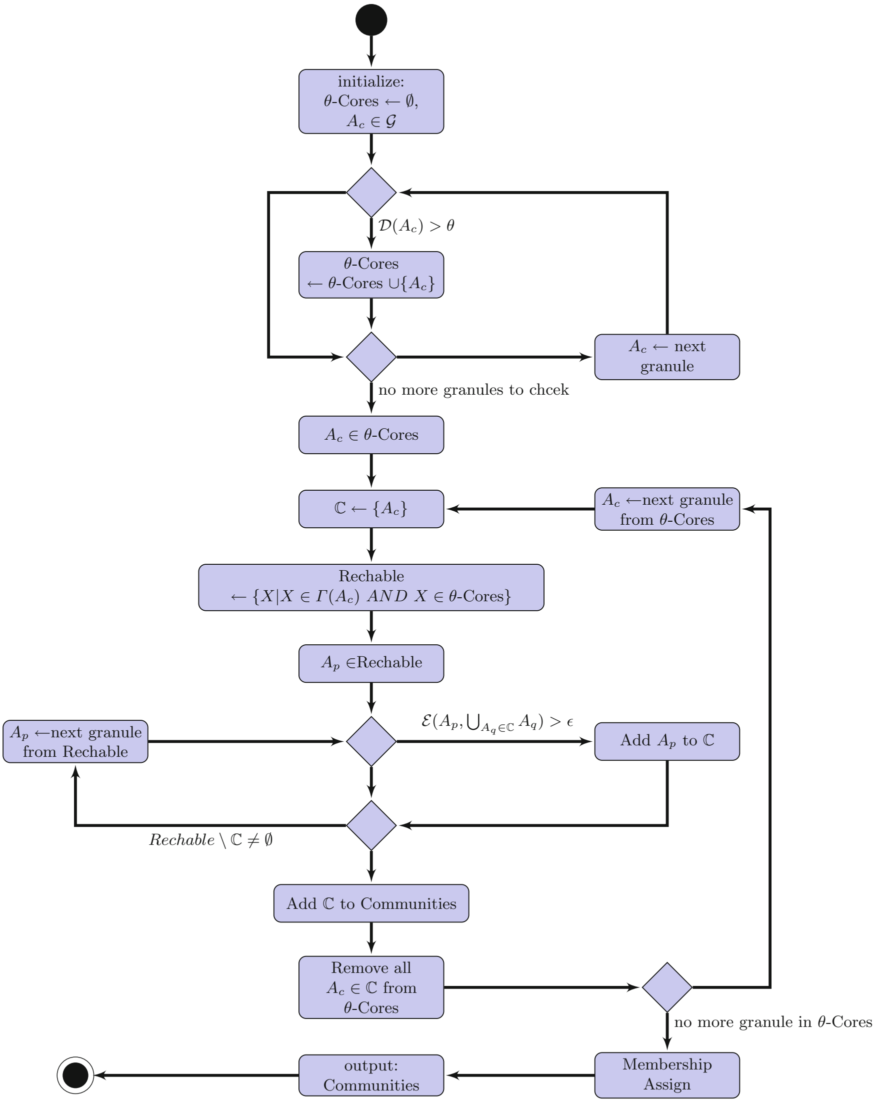

initialize:
θ-Cores ← $\emptyset$,
$A_c \in \mathcal{G}$
$\mathcal{D}(A_c) > \theta$
θ-Cores
← θ-Cores $\cup\{A_c\}$
A_c ← next granule
no more granules to chcek
$A_c \in \theta$-Cores
$\mathbb{C} \leftarrow \{A_c\}$
A_c ←next granule from θ-Cores
Rechable
← $\{X | X \in \Gamma(A_c)\ AND\ X \in \theta$-Cores$\}$
A_p ∈Rechable
$\mathcal{E}(A_p, \bigcup_{A_q \in \mathbb{C}} A_q) > \epsilon$
A_p ←next granule from Rechable
Add A_p to $\mathbb{C}$
$Rechable \setminus \mathbb{C} \neq \emptyset$
Add $\mathbb{C}$ to Communities
Remove all $A_c \in \mathbb{C}$ from θ-Cores
no more granule in θ-Cores
Membership Assign
output: Communities

Fig. 10 Block diagram of FRC-FGSN algorithm

Orphans: A node is said to be orphan if it is not a member of any identified community.

Given a social network, the algorithm (FRC-FGSN) finds its various communities (Definition 1) with fuzzy-rough description (Eq. 19) defined over the granular model (Eq. 9) of knowledge representation. Nodes not included as a part of any community are designated as orphans. A block diagram of the algorithm is shown in Fig. 10 [45].

LFR Benchmark Data: LFR benchmark data is one of the popular benchmark data for comparing community detection algorithms [48]. Later, it was modified to accommodate more properties of network and communities viz. directed, weighted network and overlapping communities [47]. The idea is to generate network graphs based on various parameters. These parameters are

- 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)

With LFR data, we compare the identified community structures with the output of three popular graph theoretic algorithms. These are, centrality based community detection method [28], Modularity optimization method [62] and k-clique percolation method (CPM) [71]. 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.

Normalized fuzzy mutual information [45] is used to compare different community detection algorithms. For two community structures $\mathbb{C}^X$ and $\mathbb{C}^Y$ the NFMI value can be calculated as

$$NFMI(\mathbb{C}^X : \mathbb{C}^Y) = \frac{1}{2}\left[\frac{H(\mathbb{C}^X) - H(\mathbb{C}^X|\mathbb{C}^Y)}{H(\mathbb{C}^X)} + \frac{H(\mathbb{C}^Y) - H(\mathbb{C}^Y|\mathbb{C}^X)}{H(\mathbb{C}^Y)}\right] \tag{20}$$

where $H(\mathbb{C}^X|\mathbb{C}^Y)$(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 $\mathbb{C}^X$). 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

$$H(\mathbb{C}^X) = -\sum_{P\in\mathbb{C}^X} \lambda_P^X \log_2(\lambda_P^X) \tag{21}$$

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

In the experiments, the size of the network is fixed to 1001 and we vary the other variables, and analyze the algorithms and their performance. The benchmark data generated by 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, for nodes in overlapping region, we are assigning different memberships for belonging to different communities, but the network generated by LFR assigns unity value to these nodes. So, it is not the perfect data set to test our algorithms, yet results are convincing, as described below.

Fig. 11 Comparative results with different values of mixing parameter. Network size: 1001; Min community size: 150; Max community Size: 250

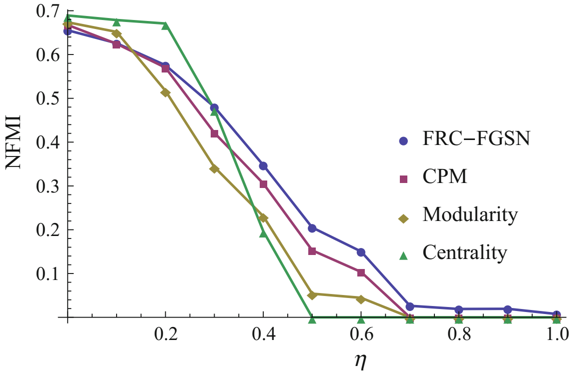
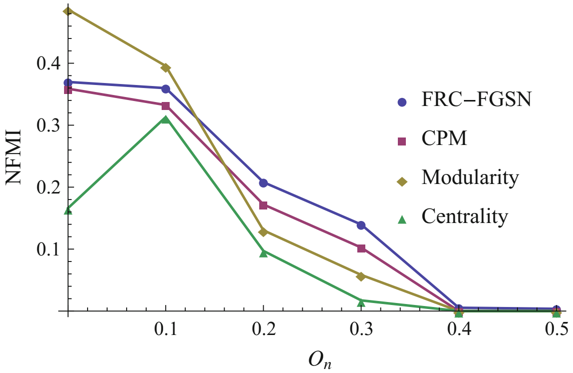


Fig. 12 Comparison showing variation of NFMI for different fraction of overlapping community. Network size: 1001; Mixing parameter: 0.4; Min community size: 150; Max community size: 250



First, we vary the mixing parameter 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. Figure 11, 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 at 0.4. Results are reported in Fig. 12. It shows that the proposed FRC-FGSN produces superior performance for O_n ranging from 0.2 to 0.4 and second best for $O_n < 0.2$.

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. 13. 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 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 decrease as we reduce the number of granules from 100

Fig. 13 Plot showing the performance on number of granules for LFR data

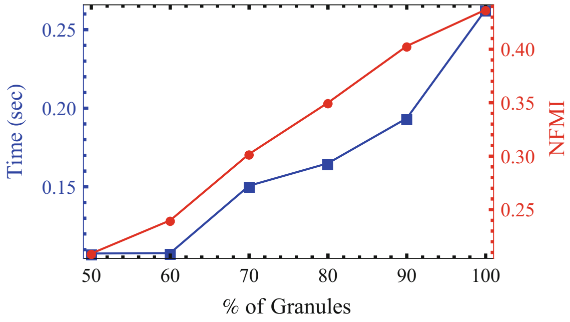


Table 3 Characteristics of generated data sets

Properties	Datasets				
	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5
Nodes	8404	16998	25761	34328	42965
Edges	163397	396178	660909	934708	1233418
Closed triangles	349156	879612	1526011	2178342	2952935
Open triangles	22228753	66714578	125632613	183297754	256587753
Approx. full diameter	4	5	5	5	5
90% effective diameter	2.875267	2.900036	2.931684	2.932677	2.935486

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 FGSN one may obtain execution benefits in the algorithm.

4.6 *Scalability of FGSN*

We conducted experiments to understand the performance of FGSN with the growing number of links in the network. We used LDBC DATAGEN [75] to generate social network data of different scale. LDBC DATAGEN is a synthetic graph data generator which internally uses S3G2 [74] algorithm to generate social network data. DATAGEN generates realistic social networks based on the link distributions found in a real social network such as Facebook [75]. DATAGEN follows the MapReduce [15] paradigm, allowing for the generation of large data sets on commodity clusters.

With the help of DATAGEN, we generated five data sets. Characteristics of these networks are listed in Table 3. We observed the time required to convert these networks into FGSN model. Python modules NetworkX and Pandas are used for graph-

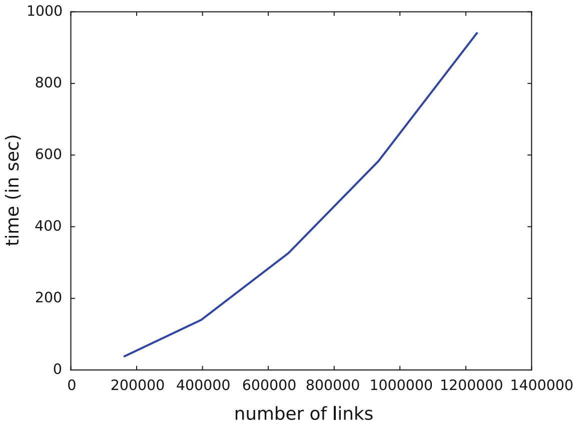


Fig. 14 Variation of conversion time with the number of links in the network

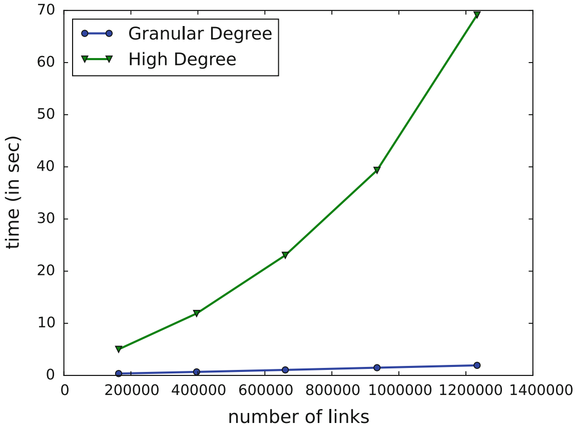


Fig. 15 Variation of execution time with number of links in the network

based operations and granule based operations respectively. Figure 14 shows the results in graphical format. The x-axis shows the number of links in the network and y-axis shows the time required for the conversion. As expected, the time increases with the number of links in the network. Clearly the curve shows a quadratic pattern. It is seen that for a network with over 1.23 millions of edges require slightly over 15 min of time for the conversion. A point to note here is that this time is required only once and the granular social network thus produced can be saved in flat files or database for future use. After the conversion, one may efficiently execute algorithms designed for the granular social networks. For example, the granular degree heuristic algorithm for the problem of target set selection is magnitude faster than the high degree heuristic. Figure 15 shows the execution time for extracting 50 seeds with aforesaid two algorithms on different data sets. The time here corresponds to 100 runs. The granular degree heuristics take only 1.93 s for the network with 1.23 millions edges, where as for the same network high degree heuristic took 69.14 s.

5 Discussions and Conclusions

In this chapter, we described a model of social network based on fuzzy granulation theory. Granules in the model characterize the closely operative groups formed within the highly overlapping neighborhood of social networks. The presence of vaguely defined closeness in relationships is modeled through the fuzzy set theory. The model is named *fuzzy granular social network* (FGSN).

In the graph representation of network, an individual node is used as an actor, whereas in FGSN, a granule is considered as an actor. A granule is constructed around each node in the network. This enables to capture the maximum information of the network inside the FGSN model. Under this granular framework, characteristics of a network are described using various measures defined over one or more granules. These measures include granular degree, granular betweenness and granular embeddedness.

The FGSN framework 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 has different roles for its different actors, then a modification may be required to accommodate such characteristics.

The data used in the experiment are collected with the view of graph representation. Hence, we had to convert such graph networks into the new knowledge representation of FGSN. Time taken for these conversions in seconds is seen to be 3.61, 12.54 and 7.09×10^3 for the Zachary karate club network, Dolphin social graph and Political blog network respectively. Once the modeling is complete, algorithms for different tasks of network analysis can be formulated.

A point to note here is that FGSN only encodes the structural information of the network. However for online social networks, many other contents (like posts, images, tags and profile) are also available to attribute with the actors. How to encode these information inside the granular social network model is not addressed in the current study.

Two major tasks concerning social network analysis are provided in this article. These are target set selection and community detection. Granular degree heuristic algorithm described for target set selection on FGSN uses granular degrees to rank the influencing nodes. Top k nodes selected from this ranked list are then used as the seed for the problem of target set selection. This selected target set is seen to perform better for most of the test cases in the undirected social networks of karate club network and Dolphin social graph. For directed network it is at par with the high degree heuristic but lower than that of the diffusion degree heuristics.

The output communities found by FRC-FGSN 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 in belonging. In the process orphan (nodes with zero membership to all communities) are detected automatically.

Normalized fuzzy mutual information (NFMI) quantifies well the goodness of the identified communities. Larger is the value of NFMI between two community structures, higher is their similarity. It is shown that the FRC-FGSN algorithm produces superior outcome as compared to other popularly known community detection algorithms when the network contains overlapped communities.

Social networks available from popular mobile and Internet applications produce data in huge scale. These data show all the characteristics of Big data. Scalability is one of the important issues for Big data analysis. In case of big social networks, FGSN has the following two advantages over the graph modeling. First, in FGSN, the network properties of a node are embedded inside the granule constructed around it. If an algorithm demands to work on fewer nodes rather than the full network then one may avoid feeding the full network into the algorithm and yet can get the network properties from the granular characteristics. Even for the global property analysis, for reducing the execution time of data processing one may restrict the number of granules either based on a threshold, decided over the cardinality of the granule, or with human intervention. Experimentally, we found that with the reduction in number of granules, the rate of improvement in execution time is exponential while the rate of drop in accuracy is linear. Second, FGSN supports asynchronous nature of distributed computing better than the graph modeling. Two of the major challenges involving the distributed computing are, (1) coping with the intrinsic asynchrony between the different entities, and (2) coping with the spatial distribution of these computing entities. Granules may be more effectively fed into such asynchronous distributed systems where one computing unit will only deal with a subset of granules. Whereas feeding a graph model in such system is difficult. With graphs an additional care needs to be taken to work with synchronous algorithms in a distributed environment.

It is seen experimentally that the algorithm scales well with the growing number of links in the network. DATAGEN, a Hadoop/MapReduce based data generator is used to generate synthetic data for scalability analysis. The growth in execution time with the number of links for granular degree heuristics is found to be linear and the slope is also very low.

The model of FGSN is seen here to perform effectively and efficiently for two of the major applications in the domain of social network analysis. There are other applications in social network analysis e.g., link prediction and evolution of social network which are also very important to study. For example existing algorithms on link prediction through graph implementation find the similarity between two nodes which in turn provides the plausibility that a link may form between them in future. Similarities can easily be identified using the normalized embeddedness measure in FGSN. If two granules are highly embedded to each other, then there is a high possibility that there would be a link between the centers of the granules in the network.

Although, some of the algorithms available in the domain might provide better solutions as compared to the proposed methodologies, the way of modeling a network with FGSN opens a new avenue and provides directions on using the established granular computing theory and other efficient data mining techniques into the demanding dynamics of social networks and related problems with a scope of newly defined measures and efficient algorithms.

Acknowledgements The work was completed while S.K. Pal held the J.C. Bose National Fellowship, Indian National Academy of Engineering Chair professorship, and DAE Raja Ramanna Fellowship.

References

1. L.A. Adamic, N. Glance, The political blogosphere and the 2004 US election: divided they blog, in *Proceedings of the 3rd International Workshop on Link Discovery, LinkKDD '05* (ACM, Chicago, 2005), pp. 36–43
2. Y.Y. Ahn, J.P. Bagrow, S. Lehmann, Link communities reveal multiscale complexity in networks. Nature **466**(7307), 761–764 (2010)
3. A. Anthony, S. Biesan, Block Modeling in Large Social Networks with Many Clusters. Technical report (2012)
4. C. Aslay, N. Barbieri, F. Bonchi, R. Baeza-Yates, Online topic-aware influence maximization. Proc. VLDB Endow. **8**(6), 666–677 (2015)
5. A.L. Barabási, R. Albert, Emergence of scaling in random networks. Science **286**(5439), 509–512 (1999)
6. S.A. Boorman, H.C. White, Social structure from multiple networks. ii. role structures social structure from multiple networks. Am. J. Sociol. **81**(6), 1384–1446 (1976)
7. D. Chakraborty, B.U. Shankar, S.K. Pal, Granulation, rough entropy and spatiotemporal moving object detection. Appl. Soft Comput. J. **13**(9), 4001–4009 (2013)
8. S. Chattopadhyay, C.A. Murthy, S.K. Pal, Fitting truncated geometric distributions in large scale real world networks. Theor. Comput. Sci. **551**, 22–38 (2014)
9. S. Chen, Y. Li, Dynamic grade on the major hazards using community detection based on genetic algorithm, in *Proceedings of 2009 International Conference on Signal Processing Systems* (IEEE, Singapore, 2009), pp. 713–717
10. W. Chen, C. Wang, Y. Wang, Scalable influence maximization for prevalent viral marketing in large-scale social networks, in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, 2010a), pp. 1029–1038
11. W. Chen, Y. Wang, S. Yang, Efficient influence maximization in social networks, in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM Press, Paris, 2009), pp. 199–208
12. W. Chen, Y. Yuan, L. Zhang, Scalable influence maximization in social networks under the linear threshold model, in *2010 IEEE International Conference on Data Mining* (IEEE, New Jersey, 2010b), pp. 88–97
13. E. Cohen, D. Delling, T. Pajor, R.E. Werneck, Sketch-based influence maximization and computation: scaling up with guarantees, in *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management (CIKM '14)* (ACM Press, New York, 2014), pp. 629–638
14. G.B. Davis, K.M. Carley, Clearing the FOG: fuzzy, overlapping groups for social networks. Soc. Netw. **30**(3), 201–212 (2008)
15. J. Dean, S. Ghemawat, mapreduce: simplified data processing on large clusters. Commun. ACM **51**(1), 107 (2008)
16. P. Domingos, M. Richardson, Mining the network value of customers, in *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, San Francisco, CA, 2001), pp. 57–66

17. D. Easley, J. Kleinberg, Networks, Crowds, and Markets: Reasoning about a Highly Connected World (Cambridge, Cambridge University Press, 2010)
18. P.a. Estevez, P. Vera, K. Saito, Selecting the most influential nodes in social networks, in *Proceedings of 2007 International Joint Conference on Neural Networks* (IEEE, New Jersey, 2007), pp. 2397–2402
19. T.S. Evans, R. Lambiotte, Line graphs of weighted networks for overlapping communities. Eur. Phys. J. B **77**(2), 265–272 (2010)
20. T. Falkowski, A. Barth, M. Spiliopoulou, DENGRAPH: a density-based community detection algorithm, in *Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence (WI'07)* (IEEE, Washington, 2007), pp. 112–115
21. 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, 1999), pp. 251–262
22. T.F. Fan, C.J. Liau, T.Y. Lin, Positional analysis in fuzzy social networks, in *Proceedings of 2007 IEEE International Conference on Granular Computing (GRC 2007)* (IEEE, Silicon Valley, 2007), pp. 423–428
23. I.J. Farkas, D. Ábel, G. Palla, T. Vicsek, Weighted network modules. New J. Phys. **9**(6), 180 (2007)
24. S. Fortunato, Community detection in graphs. Phys. Rep. **486**(3–5), 75–174 (2010)
25. O. Frank, Estimation and sampling in social network analysis, in Encyclopedia of Complexity and Systems Science, by R.A. Meyers (ed.) (Springer, New York, 2009), pp. 8213–8231
26. A. Ganivada, S. Dutta, S.K. Pal, Fuzzy rough granular neural networks, fuzzy granules, and classification. Theor. Comput. Sci. **412**(42), 5834–5853 (2011)
27. K.J. Gile, M.S. Handcock, Respondent-driven sampling: an assessment of current methodology. Sociol. Methodol. **40**(1), 285–327 (2010)
28. M. Girvan, M.E.J. Newman, Community structure in social and biological networks. Proc. Natl. Acad. Sci. U.S.A **99**(12), 7821–7826 (2002)
29. M. Gomez-Rodriguez, B. Schölkopf, Influence maximization in continuous time diffusion networks, in *Proceedings of the 29th International Conference on Machine Learning (ICML-12)* (Edinburgh, 2012), pp. 313–320
30. A. Goyal, W. Lu, L. Lakshmanan, CELF++: optimizing the greedy algorithm for influence maximization in social networks, in *Proceedings of the 20th International Conference Companion on World Wide Web* (ACM, New York, 2011), pp. 47–48
31. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edn. (Springer Series in Statistics, Springer, New York, 2009)
32. P. Hońko, Association discovery from relational data via granular computing. Inf. Sci. **234**(2), 136–149 (2013)
33. H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, aL Barabási, The large-scale organization of metabolic networks. Nature **407**(6804), 651–654 (2000)
34. G. Jiang, F. Ma, J. Shang, P.Y. Chau, Evolution of knowledge sharing behavior in social commerce: an agent-based computational approach. Inf. Sci. **278**, 250–266 (2014)
35. L.j. Kao, Y.P. Huang, Mining influential users in social network, in *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2015* (Hong Kong, 2015), pp. 1209–1214
36. S. Kelley, M. Goldberg, M. Magdon-Ismail, K. Mertsalov, A. Wallace, Defining and discovering communities in social networks, in *Handbook of Optimization in Complex Networks* (2012), pp. 139–168
37. D. Kempe, J. Kleinberg, É. Tardos, Maximizing the spread of influence through a social network, in *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM Press, New York, NY, 2003), p. 137
38. D. Kempe, J. Kleinberg, É. Tardos, Influential nodes in a diffusion model for social networks. Autom. Lang. Program. **3580**, 1127–1138 (2005)
39. Y.A. Kim, R. Phalak, A trust prediction framework in rating-based experience sharing social networks without a Web of Trust. Inf. Sci. **191**, 128–145 (2012)

40. J. Kleinberg, Cascading behavior in networks: algorithmic and economic issues, in *Algorithmic Game Theory*, by eds. N. Nisan, T. Roughgarden, E. Tardos, V.V. Vazirani (Cambridge, Cambridge University Press, 2007), pp. 613–632
41. J. Kleinberg, S. Suri, É. Tardos, T. Wexler, Strategic network formation with structural holes, in *Proceedings of the 9th ACM Conference on Electronic Commerce - EC'08* (ACM Press, New York, USA, 2008), pp. 284–293
42. B. Krishnamurthy, J. Wang, On network-aware clustering of web clients, in *Proceedings of of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication, SIGCOMM '00* (ACM, New York, Stockholm, 2000), pp. 97–110
43. L. Kuang, X. Tang, M. Yu, Y. Huang, K. Guo, A comprehensive ranking model for tweets big data in online social network. EURASIP J. Wireless Commun. Netw. **2016**(1), 46 (2016)
44. S. Kundu, S.K. Pal, FGSN: fuzzy granular social networks - model and applications. Inf. Sci. **314**, 100–117 (2015a)
45. S. Kundu, S.K. Pal, Fuzzy-rough community in social networks. Pattern Recognit. Lett. **67**(2), 145–152 (2015b)
46. M. Lahiri, M. Cebrian, The genetic algorithm as a general diffusion model for social networks, in *Proceedings of the 24th AAAI Conference on Artificial Intelligence* (Atlanta, Georgia, 2010), pp. 494–499
47. A. Lancichinetti, S. Fortunato, Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. Phys. Rev. E. **80**(1), 016118 (2009)
48. A. Lancichinetti, S. Fortunato, F. Radicchi, Benchmark graphs for testing community detection algorithms. Phys. Rev. E. **80**(1), 016118 (2008)
49. J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, N. Glance, N, Cost-effective outbreak detection in networks, in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM Press, San Jose, 2007), pp. 420–429
50. J. Leskovec, K.J. Lang, M. Mahoney, Empirical comparison of algorithms for network community detection, in *Proceedings of the 19th International Conference on World Wide Web - WWW '10* (Raleigh, 2010), p. 631
51. G. Li, S. Chen, J. Feng, K.J. Tan, W.s. Li, Efficient location-aware influence maximization, in *Proceedings of the 2014 ACM SIGMOD international conference on Management of data (SIGMOD'14)* (Snowbird, 2014), pp. 87–98
52. L. Li, S. Li, X. Chen, A new genetics-based diffusion model for social networks, in *Proceedings of 2011 International Conference on Computational Aspects of Social Networks (CASoN)* (IEEE, Salamanca, Spain, 2011), pp. 76–81
53. O. Liu, K.L. Man, W. Chong, C.O. Chan, Social network analysis and big data, in *Proceedings of the International Multi Conference of Engineers and Computer Scientists*, vol. II (Hong Kong, 2016), pp. 6–7
54. W, Liu., X, Jiang, M, Pellegrini, X, Wang, Discovering communities in complex networks by edge label propagation. Sci. Rep. **6** (2016)
55. 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), 396–405 (2003)
56. 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), 741–52 (2010)
57. F.D. Malliaros, M. Vazirgiannis, Clustering and community detection in directed networks: a survey. Phys. Rep. **533**(4), 95–142 (2013)
58. J.L. Moreno, *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)
59. P.S. Nair, S.T. Sarasamma, Data mining through fuzzy social network analysis, in *Proceedings of the 26th International Conference of North American Fuzzy Information Processing Society* (IEEE, San Diego, California, 2007), pp. 251–255

60. R. Narayanam, Y. Narahari, A Shapley value-based approach to discover influential nodes in social networks. IEEE Trans. Autom. Sci. Eng. **8**(1), 130–147 (2011)
61. M. Narayanan, A. Cherukuri, A study and analysis of recommendation systems for location-based social network (LBSN) with big data. IIMB Manag. Rev. **28**(1), 25–30 (2016)
62. M. Newman, Fast algorithm for detecting community structure in networks. Phys. Rev. E **69**(6), 066133 (2004)
63. M. Newman, M. Girvan, Finding and evaluating community structure in networks. Phys. Rev. E **69**(2), 1–15 (2004)
64. M. Newman, Modularity and community structure in networks. Proc. Natl. Acad. Sci. U.S.A **103**(23), 8577–8582 (2006)
65. G.K. Orman, V. Labatut, The effect of network realism on community detection algorithms, in *Proceedings of the 2010 International Conference on Advances in Social Networks Analysis and Mining* (IEEE, Odense, Denmark, 2010), pp. 301–305
66. S.K. Pal, Granular mining and rough-fuzzy pattern recognition: a way to natural computation. IEEE Intell. Inf. Bull. **13**(1), 3–13 (2012)
67. S.K. Pal, S. Kundu, C.A. Murthy, Centrality measures, upper bound, and influence maximization in large scale directed social networks. Fundam. Inf. **130**(3), 317–342 (2014)
68. S.K. Pal, S.K. Meher, Natural computing: a problem solving paradigm with granular information processing. Appl. Soft Comput. J. **13**(9), 3944–3955 (2013)
69. S.k. Pal, S.K. Meher, A. Skowron, Data science, big data and granular mining. Pattern Recognit. Lett. **67**(2), 109–112 (2015)
70. S.K. Pal, P. Mitra, *Pattern Recognition Algorithms for Data Mining* (CRC Press, Boca Raton, 2004)
71. G. Palla, I. Derényi, I. Farkas, T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society. Nature **435**(7043), 814–818 (2005)
72. W. Pedrycz, *Granular Computing: Analysis and Design of Intelligent Systems* (CRC Press, Boca Raton, 2013)
73. W. Pedrycz, A. Skowron, V. Kreinovich (eds.), *Handbook of granular computing* (Wiley, Sussex, 2008)
74. M.D. Pham, P. Boncz, O. Erling, S3G2: a scalable structure-correlated social graph generator, in *Selected Topics in Performance Evaluation and Benchmarking: 4th TPC Technology Conference, TPCTC 2012, Istanbul, Turkey, August 27, 2012, Revised Selected Papers*, ed. by R. Nambiar, M. Poess (Springer, Berlin, 2013), pp. 156–172
75. A. Prat, DATAGEN: Data Generation for the Social Network Benchmark (2014). http://ldbcouncil.org/blog/datagen-data-generation-social-network-benchmark
76. Y. Qin, J. Ma, S. Gao, Efficient influence maximization under TSCM: a suitable diffusion model in online social networks. Soft Comput. 1–12 (2016)
77. U. Raghavan, R. Albert, S. Kumara, Near linear time algorithm to detect community structures in large-scale networks. Phys. Rev. E **76**(3), 36106 (2007)
78. 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, London, 2002), pp. 188–200
79. F. Reid, A. McDaid, N. Hurley, Partitioning breaks communities, in *Proceedings of 2011 International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2011* (Kaohsiung City, Taiwan, 2011), pp. 102–109
80. S.A. Rice, The identification of blocs in small political bodies. Am. Polit. Sci. Rev. **21**(3), 619–627 (1927)
81. M. Richardson, P. Domingos, Mining knowledge-sharing sites for viral marketing, in *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM Press, Edmonton, Alberta, 2002), pp. 61–70
82. D. Sen, S. Pal, Generalized rough sets, entropy, and image ambiguity measures. IEEE Trans. Syst. Man Cybern. Part B **39**(1), 117–128 (2009)

83. M. Spreen, Rare Populations, Hidden Populations, and Link-Tracing Designs: What and Why? Bulletin de Méthodologie Sociologique **36**, 34–58 (1992)
84. M. Steenstrup, Cluster-based networks, in *Ad Hoc Networking, Chap*, vol. 4 (Addison-Wesley Longman Publishing Co., Inc, Boston, MA, USA, 2001), pp. 75–138
85. Z. Su, Q. Xu, Q. Qi, Big data in mobile social networks: a QoE-oriented framework. IEEE Netw. **30**(1), 52–57 (2016)
86. C. Wang, L. Deng, G. Zhou, M. Jiang, A global optimization algorithm for target set selection problems. Inf. Sci. **267**, 101–118 (2014)
87. C. Wang, W. Chen, Y. Wang, Scalable influence maximization for independent cascade model in large-scale social networks. Data Mining Knowl. Discov. **25**(3), 545–576 (2012)
88. S. Wasserman, K. Faust, *Social Network Analysis: Methods and Applications* (Cambridge University Press, Cambridge, 1994)
89. R.S. Weiss, E. Jacobson, A method for the analysis of the structure of complex organizations. Am. Sociol. Assoc. **20**(6), 661–668 (1955)
90. H.C. White, S.A. Boorman, R.L. Breiger, Social structure from multiple networks. I. Blockmodels of roles and positions. Am. J. Sociol. **81**(4), 730–780 (1976)
91. J. Yao, A.V. Vasilakos, W. Pedrycz, Granular computing: perspectives and challenges. IEEE Trans. Cybern. **43**, 1977–1989 (2013)
92. W. Zachary, An information flow model for conflict and fission in small groups. J. Anthropol. Res. **33**(4), 452–473 (1977)
93. L. Zadeh, Fuzzy sets. Inf. Control **8**, 338–353 (1965)
94. L.A. Zadeh, Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. Fuzzy Sets Syst. **90**, 111–127 (1997)
95. Y. Zeng, X. Chen, G. Cong, S. Qin, J. Tang, Y. Xiang, Maximizing influence under influence loss constraint in social networks. Expert Syst. Appl. **55**, 255–267 (2016)
96. T. Zhu, B. Wang, B. Wu, C. Zhu, Maximizing the spread of influence ranking in social networks. Inf. Sci. **278**, 535–544 (2014)