Community detection using Fire propagation and Boundary vertices algorithms

A DISSERTATION

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE AWARD OF DEGREE

OF

MASTER OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

Submitted By:

RAHUL HANOT 2K18/CSE/13

Under the supervision of

Mr. SANJAY KUMAR

(Assistant Professor)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College Of engineering)

Bawana Road, Delhi-110042

JUNE, 2020

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College Of engineering)

Bawana Road, Delhi-110042

DECLARATION

I, Rahul hanot, Roll No. 2K18/CSE/13 student of M.Tech (Compter Science & Engineering), hereby declare that the Project Dissertation titled "**Community detection using Fire propagation and Boundary vertices algorithms**" which is submitted by me to the Department of Computer Science & Engineering, Delhi Technological University, Delhi. Report of the Major II which is being submitted to Delhi Technological University, Delhi, in partial fulfillment for the requirement of the award of degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associate ship, Fellowship or other similar title or recognition.

Place: DTU, Delhi

Date: 29-05-2020

Rahul Hanot

(2K18/CSE/13)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College Of engineering)

Bawana Road, Delhi-110042

CERTIFICATE

I, hereby certify that the Project Dissertation titled *"**Community detection using Fire propagation and Boundary vertices algorithms**"* which is submitted by Rahul Hanot, Roll No. 2K18/CSE/13, Department of computer Science & Engineering, Delhi Technological University, Delhi in partial fulfillment for the requirement of the award of degree of Master of Technology (Computer Science and Engineering) is a record of a project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi

Date:

(Mr. Sanjay Kumar)
SUPERVISOR
Assistant Professor
Department of Computer Engineering Delhi
Technological University

ACKNOWLEDGEMENT

I am most thankful to my family for constantly encouraging me and giving me unconditional support while pursuing this research.

I am extremely grateful to **Mr. Sanjay Kumar** Asst. Professor, Department of Computer Science Engineering, Delhi Technological University, Delhi for providing invaluable guidance and being a constant source of inspiration throughout my research. I will always be indebted to his for the extensive support and encouragement he provided.

I also convey my heartfelt gratitude to all the research scholars of the web Research Group at Delhi Technological University, for their valuable suggestions and helpful discussions throughout the course of this research work.

Rahul Hanot

Roll-No-2K18/CSE/13

ABSTRACT

Community detection in complex networks deal with grouping related nodes together and plays a vital role to understand the functioning of the system in real-life situations. Community detection is classified as an NP-hard problem. Various algorithms are currently available for it but the problem with these existing algorithms is either they have high in time complexity or they have not able to partition the network perfectly. In this paper, we propose a novel community detection algorithm that works in two phases. In the first phase, we apply fire propagation technique in which choosing an arbitrary vertex as the core vertex and connecting an adjacent vertex to it and shapes a community this is similar to how fire spreads in real-life situations. In the second phase,we use the result of the first phase of an overlapped community and detect all boundary vertices which are belongings to more than one communities and assign them to the single community based on the weight that each core vertex assign to that particular boundary vertex using Dijkstra distance and the count of the adjacent vertex that belong that community. The proposed algorithm performs well as compared to label propagation and walk-trap algorithm in terms of modularity score using various synthetic and real-world datasets.

keywords Boundary vertices, complex network ,community detection ,core vertex ,fire propagation ,modularity score ,social network

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

The mathematical study of graphs comes under Graph theory, which is used to define the association between the objects. A graph consisting of V set of vertices that are connecting by the E set of edges. A graph can be broadly classified into two categories:

- Directed Graph
- Undirected graph

Undirected graphs are those having asymmetrical edges where Directed graph having asymmetrical edges figure 1.1 show the both directed and undirected graph. Further graph can divided on the basis of weight assoicatied A further graph can be divided on the basis of weight associate with edges in two categories:

- Weighted graph
- Unweighted graph

In the unweighted graph, the weight is associated with all edges are uniform whereas in the weighed graph each edge have different weights figure 1.2 show the both unweighted and weighted graph. In the real world, complex systems in different disciplines form networks such as metabolic networks, world wide web , protein interaction , gene regulatory network. In every field, we are using the complex network for the representation of data from sociology to from economy,from economy to information technology.

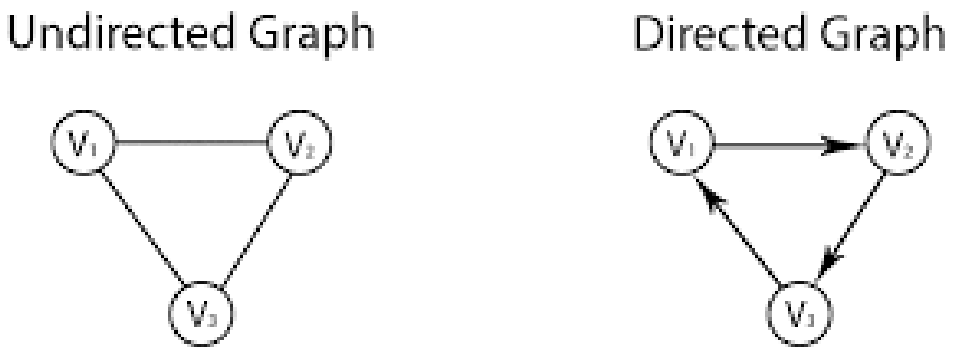

Figure 1.1 Undirected and directed graph

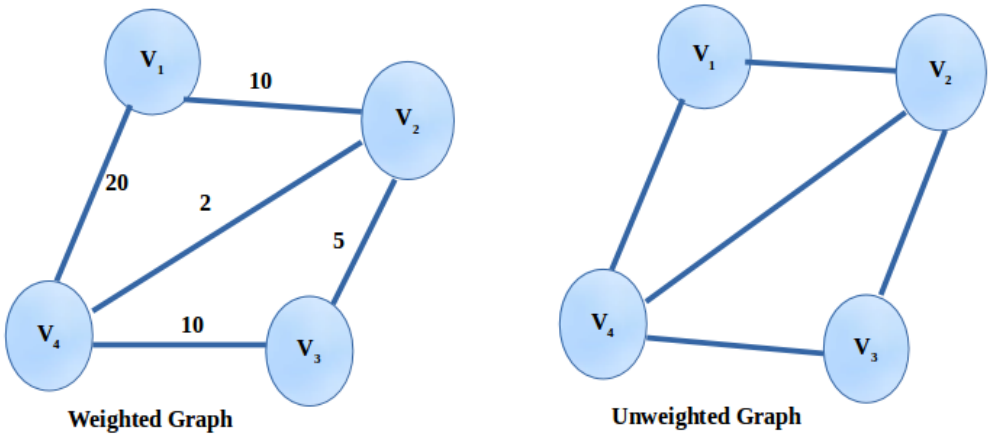

Figure 1.2 Weighted and unweighted graph

1.2 Generating graph from social network

The social network consists of serval actors such as news articles, Instagram pictures, photos, blogs, and websites, and interaction between them could be a comment over the video on youtube or the news article, tagging on the post or liking of the post. A social network can be represented in various ways among the representation using graph theory is quite popular and easy at the same time. Due to the graph's inherited structural properties, it helps us in understanding the interaction between the actor in a much better way as compared to other representations. Let the social network graph define as $G(V, E)$ where v is a set of vertices consist $v_1, v_2, v_3....v_n$ vertex represent the actors of social network and E is set of edges which represent the interaction between them as this interaction can be directed or undirected depending which social network data is used as graph representation such graph of a Facebook friend is always undirected graph where vertices are the friends and edges are the friendship between them on other hand twitter representation is always directed where vertices are a follower and the edges are the one twitter account follow the other twitter account fig 1.3 show the example the Facebook and twitter graph representation.

1.2.1 Communities in social networks

Communities in the complex network are defined as the partition or cluster of the network such that each cluster has a finite number of the vertex [5,6].that have similar properties and behavior which are different from a member of another cluster. vertices that belong to the same cluster have a large number of edges to the adjacent vertices as compared to the vertices of other clusters and also have a high ratio of intra edges to the inter edges figure 1.4 shows the process of communties detection in complex network as we can easily show in figure 1.4 vertices that are belonging to same community are large number of edges with one another as compared to the vertices to other communties. The process of the partitioning of a network into the different cluster is not the only visible in a real-world network but also in a man-made artificial network, we can see the partition of the network for example in world wide web the blogs which are based on the similar topic have are strongly connected.

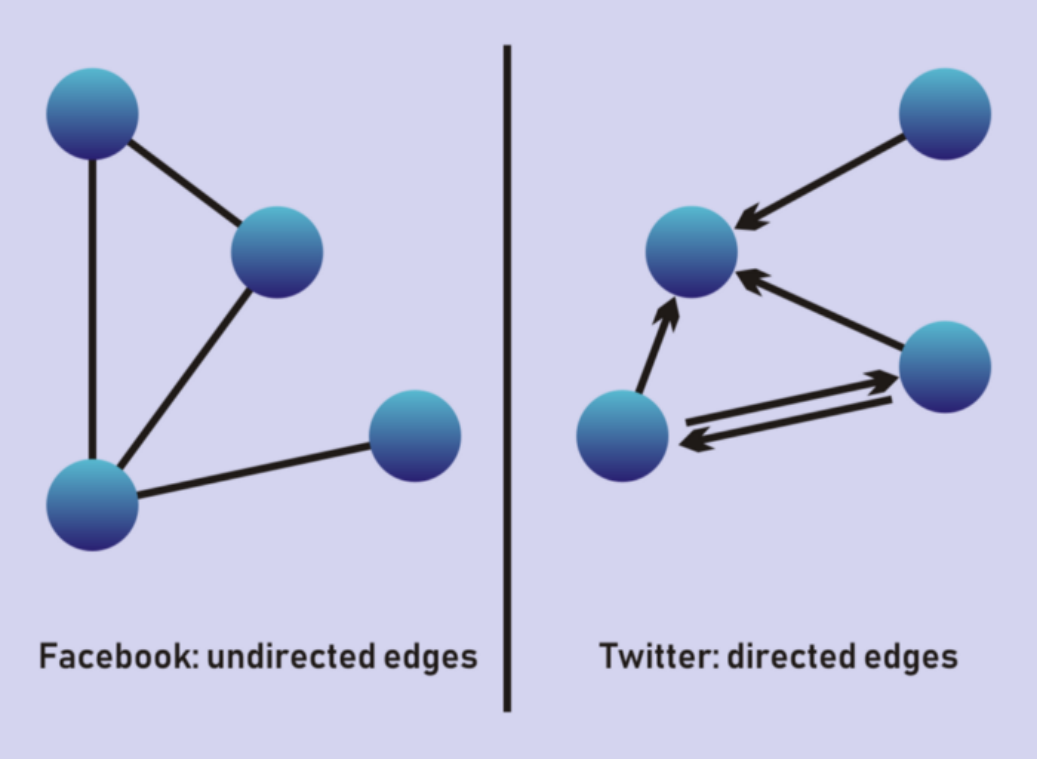

Figure 1.3 Facebook and Twitter graph representation.

1.2.2 Application of community detection

As the community detection is applicable in the [1] various field it have various application in various field as show below:

- Criminology
- Public Health
- Politics
- Customer Segmentation, Smart Advertising and Targeted Marketing
- Recommendation Systems
- Social Network Analysis
- Network Summarization and Privacy
- Link Prediction
- Community Evolution Prediction

Link Prediction

In the Link prediction process determination of future link between the vertices of a complex network, this link prediction can be used to determine the missing link , fake links, and future link. First the determination of community structure in a network then it is used to determine the expected link in the networks. ValverdeRebaza and Lopes [2] in his paper presented the novel method for link prediction and also Soundarajan and Hopcroft [3] proposed the method that increases the accuracy of similarity-based link prediction by using information obtained from the communities detected in a complex network.

Criminology

Identification of criminal user group can be done by the community detection. Those groups can be built by either real person accounts or bot accounts. terrorist-like activities and criminal ideas. In the [4] novel method is proposed for the detection of which first identify the communities in the network then

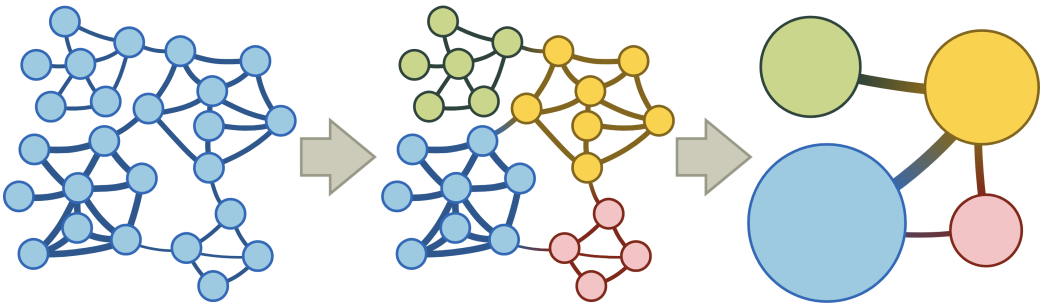

Figure 1.4 Process of communities detection

by doing manual analysis criminal activities can be detected. Another software robot named as bots used for detection of fraud follower, botnet attacks, working of bots is based on community detection in a complex network[5].

Public Health

Community detection discipline can be used for the detection of a certain group of epidemic disease. In [6] Salathe and Jones presented the effect of the community structure on dynamic diseases. Tumor and cancer can also be detected by this discipline. In [7] Bechtel et al. lung cancer detection based on community detection is presented, [8] shows the study based on a genomic dataset for detection of twelve types of cancer-based on communities detection approach. Community detection can also be used for organ detection [8,9] shows an example of organ and tissue detection.

Politics

Community detection can also be used to examine the influence of particular ideologies or individual politicians on various local groups. Community detection can be used for the detection of different types of bots [10] that can create the fake influence of the public [11].

Customer Segmentation, Smart Advertising and Targeted Marketing

Community detection can be used for smart Advertising and Targeted Marketing by the use of customer segmentation. Community detection can result in a network partition of the same interest so Companies can analyze these partitions and prepare their strategy based on and they can advertise and the market for the resulted group[12].

Network Summarization and Privacy

Community detection can be used for network summarization and privacy policies as resulted communities provide a group level point of view and summarize the network at a group level. [13] shows how community detection can be used to identify a user having multiple addresses. and also community detection used for breaking the privacy of week network like bitcoin. [14-15] shows that privacy leaked is not possible for anonymized social networks preserving the community structure of expect for the week communities.

Recommendation Systems

Community detection can be used for the recommendation system which is another important service for everybody's consumption. As we know that community detection segregates the people based on their likes due to this we can recommend people of the same liking the same thing [16-17] gives the example of a recommendation system based on community detection.

Community Evolution Prediction

From community detection, we can also predict the future form of community-based on its past and present form depending upon various events such as shrinking, merging, growing solving, etc.as this one of the hot topics in the community analysis field.[18] shows how sequential ordinary classifiers can e used for prediction whereas [19] show how this can be done as a supervised learning task.
Table 1.1 shows the applications of community detection and whether static or dynamic community method is used. Table 1.2 show the different detection method and their drawbacks

Table 1.1 Area of application and community detection method used

Area of Application	**Method of Community detection**
Criminal Identification	Static
Fraud Detection	Static
Bot Detection	Static
Criminal Activities Detection	Static
Astroturfing	Static
Customer Segmentation	Static
Community detection	Static and Dynamic
Group Segmentation	Static
Link Prediction	Static and Dynamic
Customer Evolution	Dynamic
Tissue and Organ detection	Static
Tumor or Cancer Detection	Static
Dynamic of Epidemic Spreading	Dynamic
Evolution of Influence	Dynamic

Table 1.2 Different method of community detection and their natures and their drawbacks

Method name	**Nature**	**Drawback**
Spectral Methods	Static	Not reliable for sparse method and also not computationally efficient
Statistical methods	Static	Specification of community number in advance and model selection and Time complexity is very high
Optimization based method	Static	Instability Problem and problem of resolution
Dynamic based method	Dynamic	Poor stability issue and sometimes need other clustering methods
Community detection by dependent	Dynamic and Static	Parallization is not possible and also direct community detection methods are not applicable
Simultaneous community detection on all snapshot	Dyanmic	Splitting and merging cause problem and network structure updation may cause problem

Chapter 2

Related Work

The chapter explains various kinds of Community detection algorithms as shown below:

2.1 Community detection using Genetic Algorithm

In this section generational genetic algorithm (GGA+) [20] is discussed that is governed by the modularity index to find the communities in a complex network. In this method, community count is defined initially to reduce the search space of the problem and improve the performance of the algorithm. This (GGA+) is based on safe initialization and balanced initialization. Community density is used to prevent the formation of unbalanced communities, it is used to balance the load between the communities. The size of communities is the function of the community count (total number of communities). Those sizes are employed to initialize each element of the network density vector with the number of nodes in every network.
The safe initialization is used in this method(GGA+) such that in communities every node i is connected to its adjacent neighbors of the original graph this is used to prevent undesired communities formation of disconnected nodes. Safe and balanced initialization are used to improve the performance of the algorithm. In GGA+ algorithm for the balanced initialization size of each community is computed then for the community count and safe initialization process begins by selecting the random node n_i by joining its adjacent vertex n_{ij} until the maximum size of the community is attained

and if the community is not completed with n_i then another node n_{i+1} is selected if it's not already part of any community and this process continues. As shown in figure 2.1 size of community is defined initially and size are size are community are calculated accordingly based on balanced initialization and only neighbouring node are joined to form a community to follow safe initialization.
The migration vector is initialized in this algorithm between the boundaries of the communities as shown in figure 2.3 which show the destination community of every node that could migrate. For every node,the community that has a maximum link to it then that node could migrate that community.
The crossover operation is based on the exchange of node between the communities the first node for exchange is selected as a node which has maximum to link to some other community and selection of the second node for migration is the best node that has maximum links to the community of the first node.As shown in fig 2.4 node 3 is exchange with node 5 which is best for second node as it has 1 link with destination community of node 3 so they exchange the shown in fig 2.4
The mutation operation is the process of migration (no exchange) of a node from one community to another community depending on a formerly described reproduction ratio, a good way to replicate the number of genes that could migrate in this operation. The nodes that are select for mutation migrate to the community are governed by the migration vector they destination community after the migration is may or may not be same as the node could be already in its preferable community.As shown in fig 2.5 node 2 and node 5 are selected randomly form crossover operation and they migrated according to migration vector and node 2 is migrated to its initially community only.

2.2 Community detection using edge degree betweenness centrality

This method is also an NMI based method and this method is similar to The Girvan-Newman's method (GN) greedy, hierarchical, divisive method, one of the most popular algorithms for community detection which is widely used in recent year method but unlike Girvan-Newman algorithm, this method can remove more than one edges of high degree centrality (edge betweenness) as result this algorithm is fast compare to Girvan-Newman and results in the

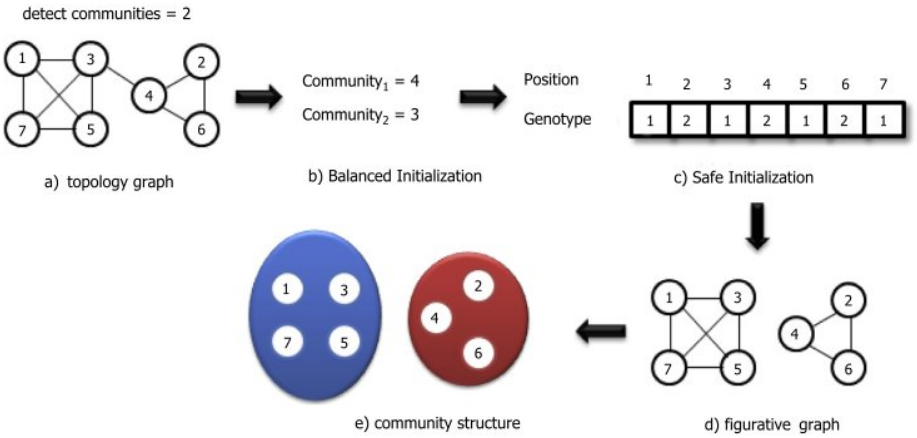

Figure 2.1 Balanced initialization and safe initialization of graph

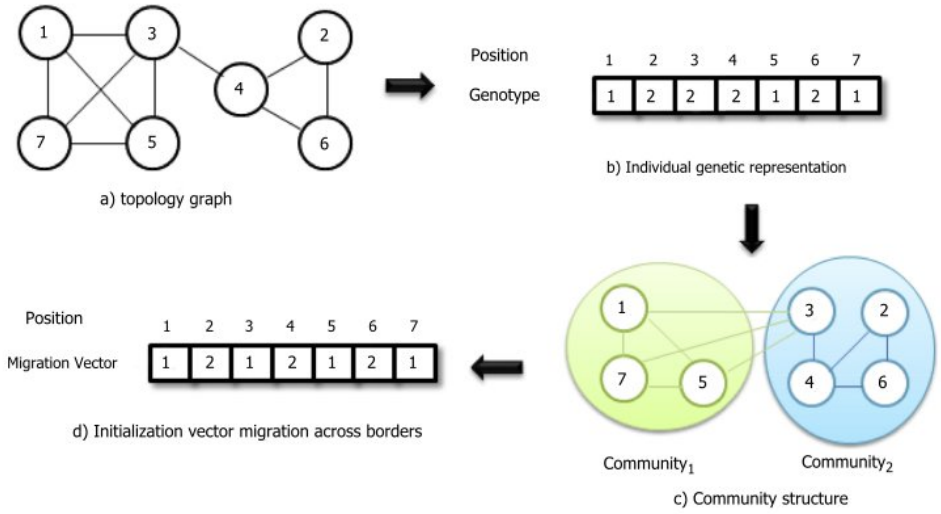

Figure 2.2: Migration vector of graph

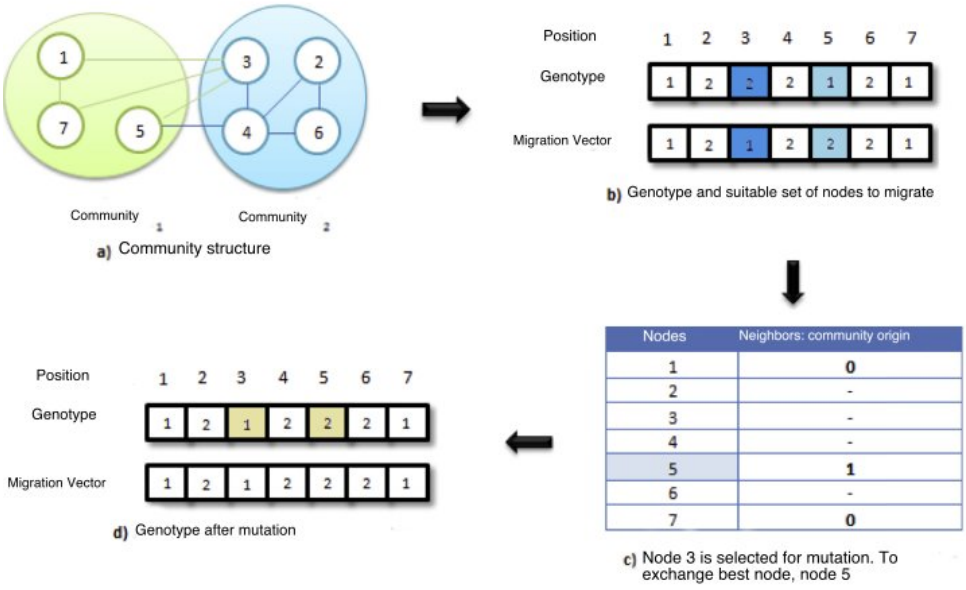

Figure 2.3: Crossover operation based on exchange of vertex between the boundaries

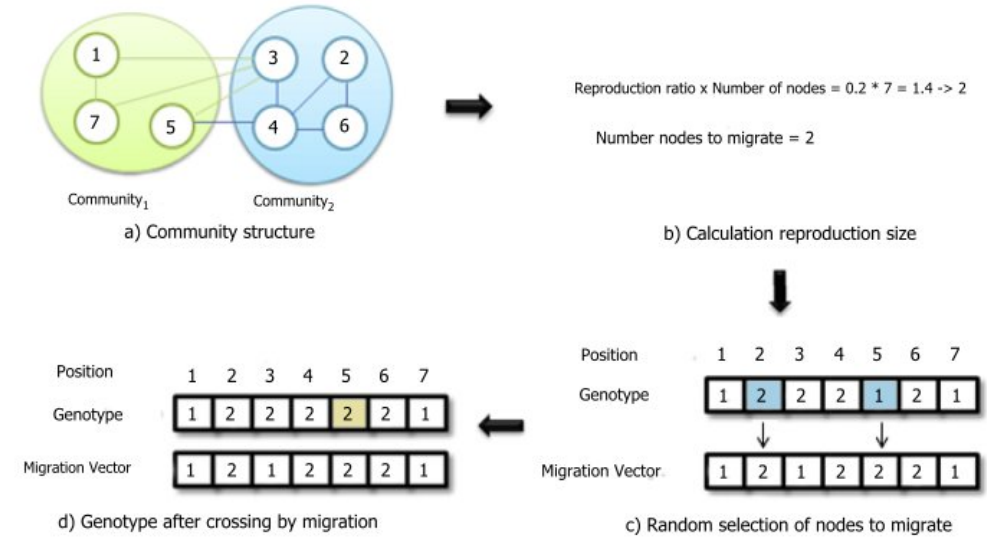

Figure 2.4: Mutations operation based on migration of vertex between the boundaries

non-overlapping community.
This algorithm has two structure as follow:

- Computations of edges degree of the network
- Removal of the most central edges from the graph

Process of edge degree computations can be done using 2 different methods:
:

- Method of edges direct adjacent vertex
- Method of line graph

In the edges direct adjacent method let the graph G having edge set E, the edge directed neighbors are the edges that satisfy the

$$D(E) = \forall(i,j) \in E|(i = u \wedge j \neq v) \vee (i \neq u \wedge j = v) \vee (i = v \wedge j \neq u) \vee (i \neq v \wedge j = u) \tag{2.1}$$

D(E) represents an edge set that has a direct relation with edges ((u, v)). For the weighted and unweighted graph the degree can be calculated via 3.3 and 3.4

$$Deg(u,v) = |D(E)| \tag{2.2}$$

$$Deg(u,v) = \sum_{(i,j)}^{|D(E)|} w(u,v) \tag{2.3}$$

The second method for calculating edges degree is the Line graph method. Let the graph $G(V, E)$ having V is set of vertices and E is set of edges the $L(G) = (E, L)$ is a line graph and two the vertex are connected in the line graph if they are connected in the introductory complex network . the line graph is also known as a dual graph or edge-to-vertex dual [21] In the line graph method vertex betweenness is based upon the shortest path, unlike the traditional method of edges betweenness [22] Hence, Line graph edges degrees, are used to calculate betweenness centrality in the graph.

Once the detection of central nodes is done using the line graph and edge neighbor method they can be removed in the next stage to increase the modularity of the network. There are various modularity methods are

define like performance, coverage, etc but for evaluation of the quality of communities detected by this method, Girvan Newman modularity method is used. The edges of the most central nodes are removing in each iteration such that after every edges removal modularity score of graph increase as result in a good network partition. The external and internal density of a cluster is defined as $\delta_{int}(C)$ and $\delta_{ext}(C)$ and the main objective of this method is to maximize the $\delta_{int}(C) - \delta_{ext}(C)$ and to increase the speed of partition method more than one edges are removed in this method in each iteration.The step algorithm is defined as below:

$$\delta_{ext}(C) = \frac{Internaledgesof C}{\frac{n_c(n_c-1)}{2}} \tag{2.4}$$

$$\delta_{ext}(C) = \frac{Externaledgesof C}{n_c(n-n_c)} \tag{2.5}$$

Algorithm 1 Community detection using fast edges betweeness

Data: Graph(V,E)
Result: List of non overlapping community C as $c_i, c_2, c_3...., c_n$
initialization ;
counter=E /* Number of edges*/
$\text{modularity}_{before} = 0$
$modularity_{after} = modularity(G)$ /* Modularity calculation */
while *(counter > 0)and* $(modularity_{before} < modularity_{after})$ **do**
    $list_{EB} = edgesBetweeness(G)$
    $list_{central} = mostCentral(G, list_{EB})$
    removeCentralEdges(G,$list_{central}$
    $modularity_{before} = modularity_{after}$
    $modularity_{after} = modularity(G)$
    if $modularity_{after} > modularity_{before}$ **then**
        C= componet(G)
    end
end
return C

2.3 Community detection algorithm based on internal force between nodes

It is a local community detection algorithm based on the seed extension, unlike the global community detection algorithm it can detect the communities without knowing the internal structure of the complex network. A local community detection algorithm excels in efficiency and effectiveness but it has quality and stability deficiencies in overlapping community detection. This algorithm has 2 stages as shown:

- Selection of seed nodes
- Community extension using seed nodes

Jaccard coefficient [23] compared with the threshold value and if the value of normalized Jaccard coefficient [24] is greater than the threshold value then that node will be used as for community extension phase. In the community extension phase first step is to sort the seed set based on the degree of the seed node and select the seed node having the highest degree from the seed set if it is not empty otherwise randomly select the node and then for each neighboring nodes calculate the fitness function for the community using 3.15 and 3.16 where f_{in} and f_{ext} is the internal and external degree of community and α is resolution parameter used for controlling the community size. Top k vertex having the highest value of fitness function calculated via [23] and [24] are used to find the fitness function with internal forces between nodes [24] and [25], and select the node which is having the highest fitness is selected to be part of a community.

$$J(u,v) = \frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)| + 1} \tag{2.6}$$

$J(u,v)$ is Jaccard coefficient between two vertex which shows the similarity between the 2 vertex and $N(u$ is set of neighbor node of vertex u.

$$F_c = \frac{f_{in}^c}{(f_{in}^c + f_{out}^c)^\alpha} \tag{2.7}$$

$$F_c^v = F_{c \cup v} - F_c \tag{2.8}$$

Where F_c^v is used to show whether the edition of vertex v makes the community more firm or not if the value of F_c^v is positive then the addition of the

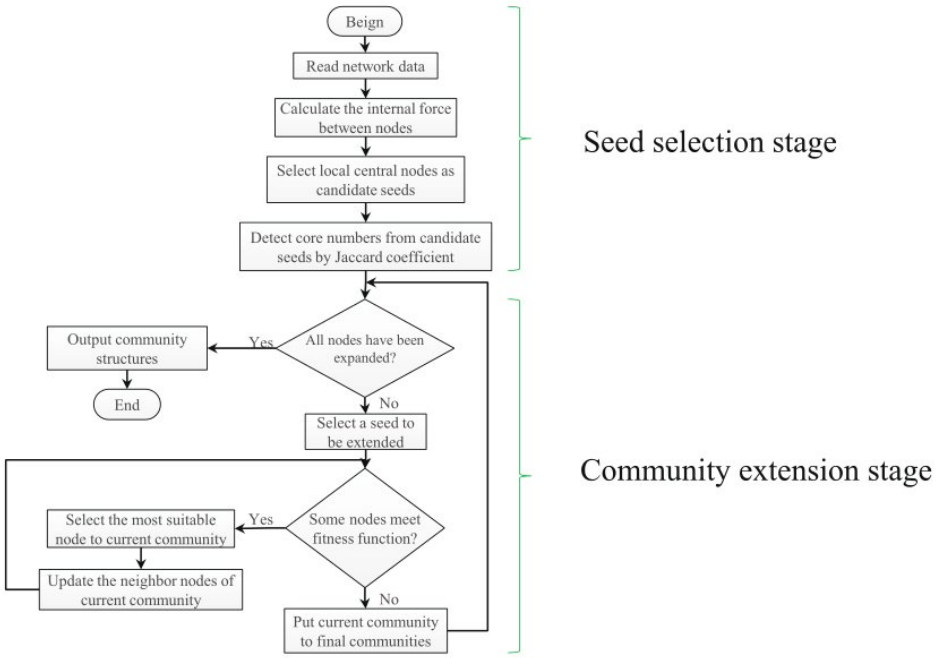

Figure 2.5: Flow graph community detection algorithm based on internal force between nodes

vertex will make a community more stable otherwise it makes community week.

$$F_G = \frac{f_{in}^{'G}}{(f_{in}^{'G} + f_{out}^{'G})^{\alpha}} \tag{2.9}$$

where the F_G represent fitness function with internl force is used to measure the tightness of community $f_{in}^{'G}$ and $f_{out}^{'G}$ are total internal and external force of the community. The flow graph of the algorithm is shown below fig 2.5:

Chapter 3

Methodology

This section presented the community detection based on a nature-inspired algorithm an extended with boundary vertex concept, As this work in 2 different stages in the first stage, we used the fire propagation method as inspired by how fire fan out in forest, and then we used a boundary vertex concept to remove overlapping community.

3.1 First stage of proposed method

The first stage of Community detection methods in a complex network is based on the idea of how the fire fan out in the forest [26] from a source object and continues to spread to its adjacent object. Using this idea we select random vertex of a complex network as seed vertex and start the formation of the community by joining its adjacent vertex.
The probability of seed vertex of catching fire is P_{seed} is 1 and an adjacent vertex to a seed vertex can catch fire which depends on the heat transmission coefficient T_h between the source and its adjacent vertices. Hence, it can be assumed that the probability of catching fire P_i near a source is directly proportional to the heat transmission coefficient T_h. Moreover, a vertex whose most of neighbors caught in a fire has greater chances of catching the fire. Let vertex v_i have n adjacent vertex caught fire then probability P_i of vertex v_i caught in fire is given by

$$P \propto T_h \tag{3.1}$$

Therefore, $P = KT_h$, where K is a constant.

$$P_i = P_{i|1}+P_{i|2}+P_{i|3}+.....P_{i|n} \tag{3.2}$$

where $P_{i|1}$ $P_{i|2}$ $P_{i|n}$ are chances of catching a fireplace for i_{th} because of adjacent vertices.
The fire catching probability at point i is denoted as P_i and it depends on fire probability $P_{i|1} \forall j \in neighbour(i)$
We can combine 3.1 and 3.2 rewrite them as 3.12:

$$P_i = \sum_{\forall i \in g.neighbour()} P_{i|j} = \sum_{\forall j \in g.neighbour()} P_j \times KT_{hj} \tag{3.3}$$

For every vertex, the threshold is defined which decides whether the concerned node catches fire or not. It is different for each vertex. Binary function is used to decide whether a vertex catches fire based on threshold value is defined as

$$M(i) = \begin{cases} 1, & \text{if } P_i \geq \text{Threshold}_i. \\ 0, & \text{if } P_i < Threshold_i. \end{cases} \tag{3.4}$$

where $M(i)$ is membership of a vertex that it belongs to any community is is greater than a calculated threshold then it vertex i will be a member of particular community else not a member of the community.
Threshold value calculation The predicted threshold of a vertex, belonging to the particular network, depends on p_{avg} and p_{avg} is calculated via 3.14

$$p_{avg} = \frac{1}{\sum_{\forall i \in g.nodes()} G.degree(i)} \tag{3.5}$$

Where G is graph and g is subgraph generated for G using two radii adjacent subgraph.
Expected number of vertices belonging to any community is v' and number of vertices that are directly connected to core vertices are v'_1 and total number of vertices that are connected to a adjacent vertex of core vertex are v'_2 these values are calculated by 3.15 and 3.16

$$v'_1 = P_1 \times v' \tag{3.6}$$

$$v'_2 = P_2 \times v' \tag{3.7}$$

P_1 and P_2 are from [26] we have seen that every vertex that is part of the subgraph may or may not finally belong to that community. Let g be a 2-radii adjacent subgraph generated from G with v number of vertices and e edges. As we know that if a vertex is a part of subgraphs that is not always true that vertex is always part of the community using this intuition we have derived $multifactor_1$ via 3.17 It provides us a count of vertex of the subgraph that is actually part of the community.

$$mf_1 = \frac{E'}{E' + 2 \times e} \tag{3.8}$$

where E' is the ratio of the total number of edges in subgraph with which any a vertex is connected to the total number of edges lie outside of the subgraph with which any vertex is connected .we can use $mf_1$to find the value of v' as shown in 3.9

$$v' = mf_1 \times v \tag{3.9}$$

In a community, any vertex is belonging to that community if most of the edges of vertex are belonging to that community or in other words we can say that higher value degree for a vertex has a higher tendency be part of the community as contrast to a smaller degree value vertex and also we gives more precedence to a vertex which has a smaller value of a degree and most of the edges belonging to a community then to a vertex which has a larger value of a degree and most of its are not lie out the community using these ideas we can define via 3.10

$$mf_2 = \frac{G.degree(i)}{g.degree(i)} \tag{3.10}$$

Where G is a graph and is g 2-radii adjacent subgraph generated from G

To find membership value defined in 3.13 for each vertex we have to calculate the threshold value shown in 3.11

$$Threshold_i = ((v'_1 \times p_{avg} + v'_2 \times p_{avg}^2) \times p_{avg}) \times \frac{2 \times e}{v^2} \times mf_2 \tag{3.11}$$

The threshold value depends upon the degree of a vertex. A vertex that has a lower value of a degree can be part of the community but if the vertex has a higher value of a degree may have less chance to be part of a community

because it may belong to other communities. The binary function defined in is also known as community membership value resembles with fire spreading phenomena. A vertex that catches fire due to seed vertex belongs to the community of seed vertex and that vertex can also catch the fire due to another vertex also so this method result in overlapping community. Algorithm of first of stage is shown below fire propagation is described below :

Algorithm 2 First stage of proposed method

```
Data: Graph(V,E)
Result: List of non overlapping community
initialization ;
 Core=[] /* List to store core node of community
 List2=[] /* List to store overlapping communities*/
 Core=[] /* List of core node while i ∈ V do
    if i ∉ List then
        Core.append(i)
         g=getNeighbourhoodSubgraph(i) /*find 2 radii adjacent subgraph */
         while j ∈ g.node() do
            calculate Pj for each vertex using Equation 3.2
             if Pj≥Threshold(G, j, g) then
                Assign vertex j to community originated from i using Equation
                 3.4
            else
                vertex j does not belong to community originated from vertex i
            end
        end
    else
        vertex i is already traversed
    end
end
```

3.2 Second stage of proposed method

Once the first stage of proposed method is over we have overlapping communities then in the next phase we remove the overlapping node which are belonging to more then one communities we assign those node into only unique community such that their is no node that is belonging to more then

one community so in the second stage of this proposed method propagation(LBN) [27]. we use the idea of label propagaiton method It reduces the update randomness in label propagation, which leads to the lack of stability in a community network. In this method, the first step is to identify all core nodes, to identify core vertex we traversed all the vertices. As we know that each node has a different impact on the network and core nodes are most important of network and they are also known as a basic node. The core node not only has a single adjacent node but also has several common adjacent nodes. The number of communities in a network is equal to the number of core vertices are identified.

$$S_i = \frac{\sum_{\forall j \in g.neighbour(i)} v_j}{d_i} \tag{3.12}$$

$$F(i) = \begin{cases} 1, & \text{if } S_i = 1. \\ 0, & \text{if } S_i \neq 1 \end{cases} \tag{3.13}$$

In 3.14 S_i represent represents the core degree of one node. J is the set of adjacent vertex of i, v_j represents the total number of common adjacent nodes of node i , and d_i is the degree of node i .$F(i)$ in 3.15 is used to identify the core vertex is value of $F(i)$ is zero then vertex is identified as core vertex and else not identified as vertex. Boundary node are those vertex which are present between the two different communities. To find the list of boundary vertex 3.14 is used.

$$g(i) = \sum L_{adj(i)} \tag{3.14}$$

where $\sum L_{adj(i)}$ represent the of list of label of adjacent vertex of vertex i if $g(i)$ is greater then one then vertex is classified as boundary vertex else not classified as boundary vertex. After identified boundary vertices, assign them to final communities. For this purpose calculate the weight that each core node assigns to boundary vertex. These weights indicate how many potential communities a boundary node has and boundary vertex assigns to the core node community that provides maximum weight to it.The calculation of weight that each core vertex assign to a boundary vertex is given by 3.17

$$w_a(x) = w^{d_{min}} \tag{3.15}$$

$$w_b(x) = \sum_{\forall i \in g.nodes()} L_i(j) \tag{3.16}$$

$$w_i(x) = w_a(x) + w_b(x) \tag{3.17}$$

d_{min} represents the minimum Dijkstra distance between core vertex and boundary vertex. w is 0.5 because which means that if only one edge is added in each iteration the basic weight will be halved. d_{xi} is a count of the adjacent vertices that belong to core vertex and d_{sum} is a degree of boundary vertex. $w_i(x)$ is the final weight that each core vertex assigns to boundary vertex.

Algorithm 3 Second stage of community detection method

Data: Graph(V,E)
Result: List of overlapping community
initialization ;

Bound=[] /* List to store Boundary communities*/

while $i \in Core$ **do**
  if $g(i) \geq 1$ **then**
    Bound.append(i) / using Equation 7 and calculate the value of g(i) and identify the boundary node*/*
  else
    vertex i is is not a boundary node
  end
end
while $i \in Bound$ **do**
  calculate the weights for each vertex i using 4.1 and assign the final the community for vertex i
end

The Proposed method is summarised as first we select the random node as shown in figure 3.1, vertex 3 and vertex 9 are selected as seed vertices for the fire propagation method these 2 vertices as shown in figure 3.2 are added to the core list then the fire fan-out process is started figure 3.3 show the result of fire fan-out process then the second stage begins as first we find the boundary vertices that belong to more than one vertex as shown in the figure 3.4 vertex 9 is boundary vertex then using Dijkstra algorithm and adjacent vertex count is used to assign the vertex 9 to the unique community as shown in fig 12 vertex 9 is assigned to the community where 3 is seed vertex .

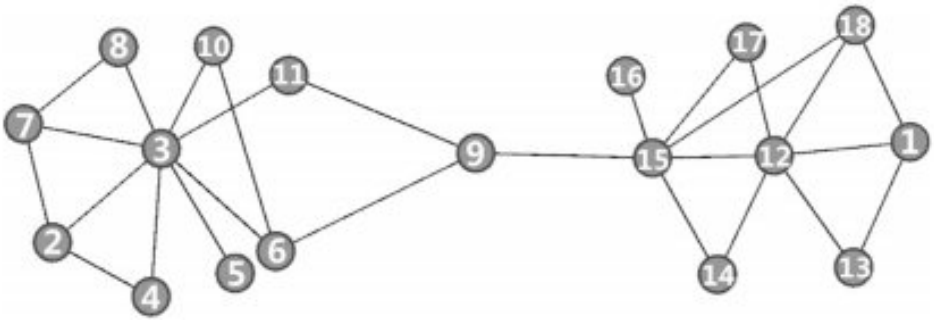

Figure 3.1: Complex network of graph

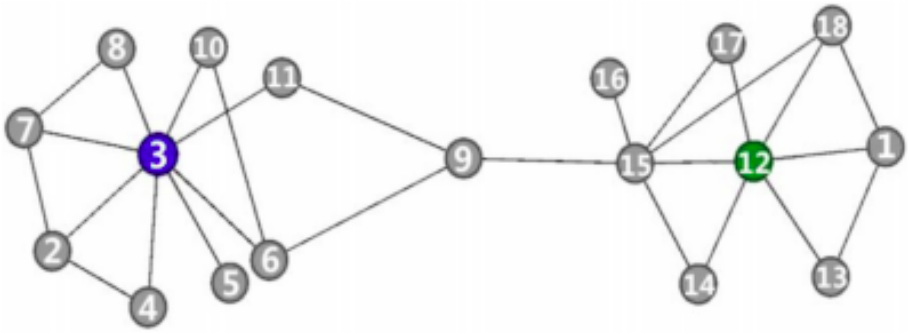

Figure 3.2: Core vertex graph

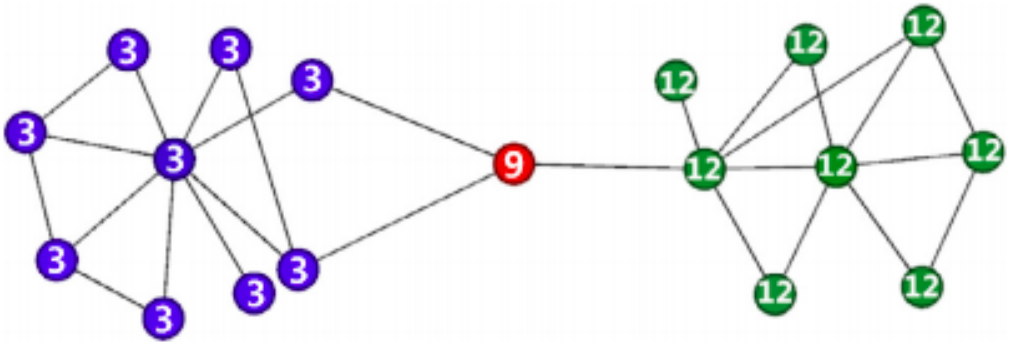

Figure 3.3: Boundary vertex graph

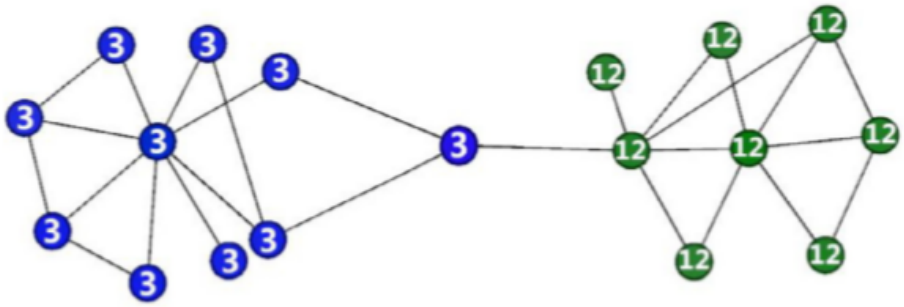

Figure 3.4: Network partition graph

Chapter 4

Experiment and Results

4.1 Datasets representation

To determine the performance of the proposed novel method we have used various different datasets that are available. We have used four real-valued datasets and four synthetic datasets which are from various categories and of different sizes, these datasets are publicly available over the internet also.

1. Jazz Musicians: It is a graph that shows the relation between jazz musicians where vertices are the jazz musician and the relation between them as edges as they played together in the same band [28].

2. Zachary karate club: It is the graph that shows the relation between karate club members where each vertex is a representation of karate clubs and the edge is the interaction between them [29].

3. Bottlenose Dolphins network: It is the graph that shows the association between them dolphins as each vertex is dolphin and the edge is the association between those dolphins [30].

4. American Political Book network: It is the graph where each vertex is a political book published around the year 2004 and the edge between the books are authors that are purchasing the many books [31].

5. Lancichinetti Fortunato Radicchi benchmark graph: The LFR benchmark algorithm is used to generate the LFR benchmark graph where the community size , community count, and every information about

the communities are known before and the degree of the graph is calculated by the power-law with a different exponent [32].

6. Gaussian random partition graph: It is a synthetic network that is created by creating k partitions, each of which is plotted with the mean s and variance s / v drawn from a normal distribution with each vertex connected to the cluster with p_{out}

7. Relaxed caveman graph: It is a synthetic graph formed by joining where one edge of individual fragments is torn down and merged with the other side in the middle of a circle of magnitude [33].

Table 4.1 Basic statistical information about datasets used in the experiment

Name of dataset	Vertex count	Edge count
Jazz musicians network	198	2742
Zachary karate club network	34	78
Bottlenose dolphin network	62	159
US political book network	105	401

By the Table 4.1 and Table 4.2 can easily point out that the data used for experimental purposes (jazz artists, American political literature database, and Zachary Karate Club information) is of different sizes and variants such as jazz artists, American political information and Zachary Karate Club information. Refers to social networking. The Bottlenose Dolphin Network refers to animal information information, the American political literature database is a database of books purchased by various authors and an artificial network of thousands of nodes [37-38].

4.2 Results and Evaluations

4.2.1 Experimental results

In this section, we have shown some of the pictorial representation of the result generated from our proposed method figure 4.1 show the barbell graph having thirteen vertices and figure 4.2 and 4.3 show the result generated from

Table 4.2: Basic statistical information about synthetic datasets used in the experiment

Name of dataset	Vertex count	specification of network
LFR benchmark graph	2600	maximum degree=25,average degree=10,minimum degree=10 ,minimum community size =10,degree power law exponent=3.0, Community power law exponent =1.1
Relaxed caveman graph	2000	size of group=10 ,number of group=20
Gaussian random partition graph	1000	size of group=30 , inter cluster probability=0.05,intra cluster probability=0.5

the proposed method and similarly figure 4.4 ,4.5 ,4.6 and 4.7 show the result of Zachary karate club

4.2.2 Performance Analysis using Modularity score

To check the performance of our proposed method and how to it perform on the various dataset we have used the various dataset size are different in sizes and types likes (US political dataset, Zachary-karate-Club and Jazz Musician), Synthetic dataset (LFR and Relaxed cavemen and Gaussian random graph) and Mammal datasets (Bootelnose dolphin dataset). As the evaluation parameter, we have used modularity which shows the accuracy of the proposed method, and to determine how well our algorithm is performed we have compared the modularity score [34] given by the proposed method with the Label propagation and walked trap algorithms. The table shows the result that is generated for the above experiment along with the modularity score of comparative methods.

$$Q = \frac{1}{2m} \sum_{vw} [A_{vw} - \frac{K_v K_w}{2m}] \frac{S_v Sw}{2} \tag{4.1}$$

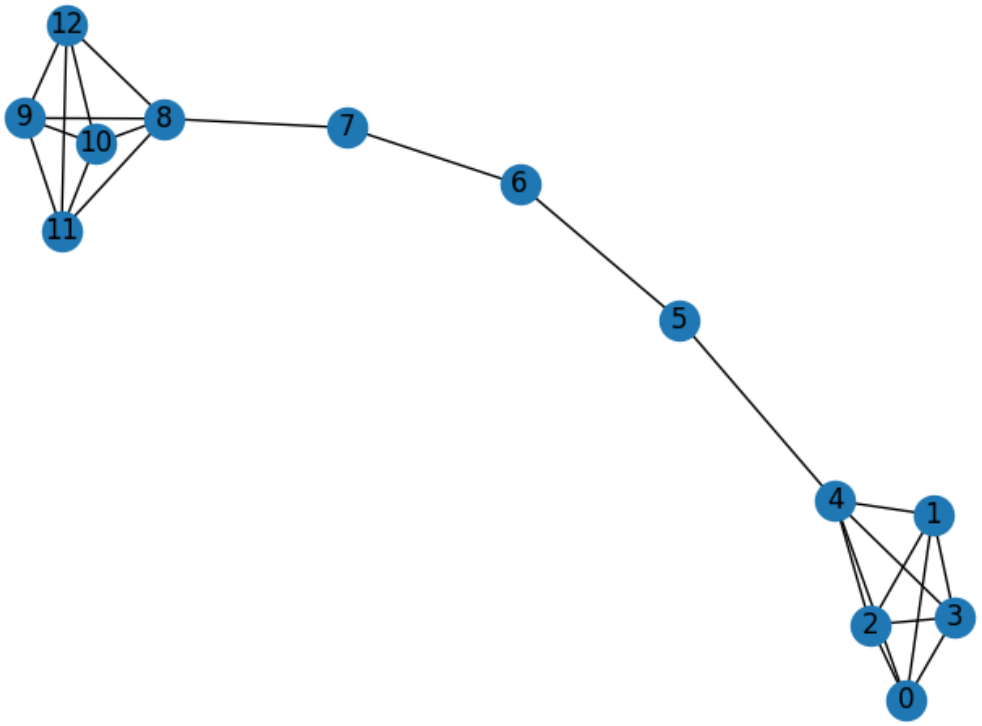

Figure 4.1: Barbell graph (5,3)

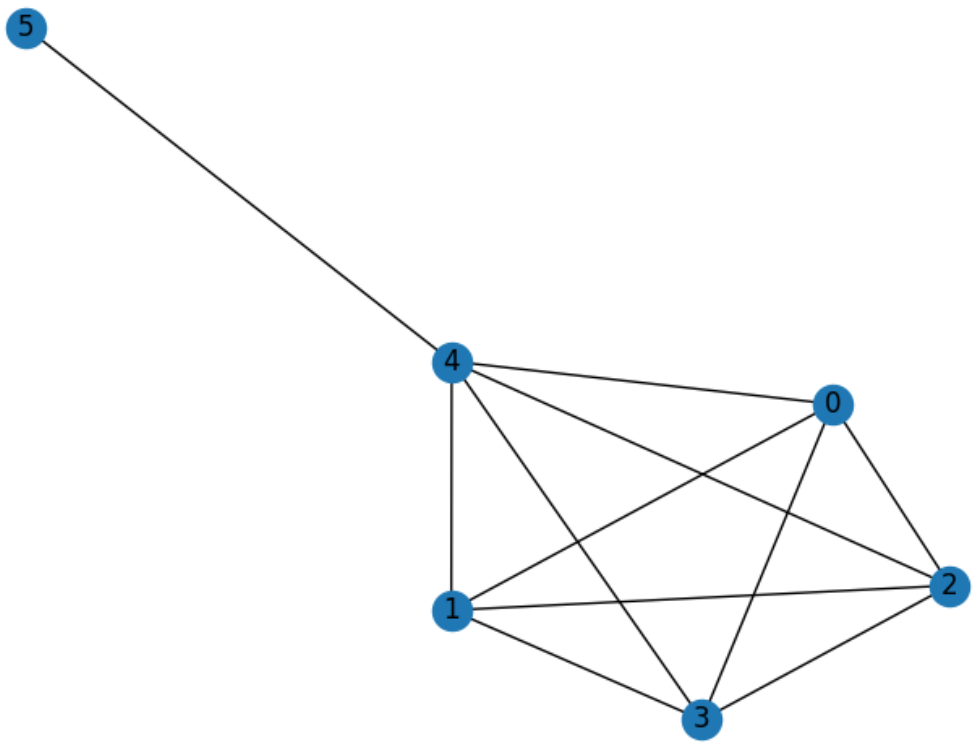

Figure 4.2: First community

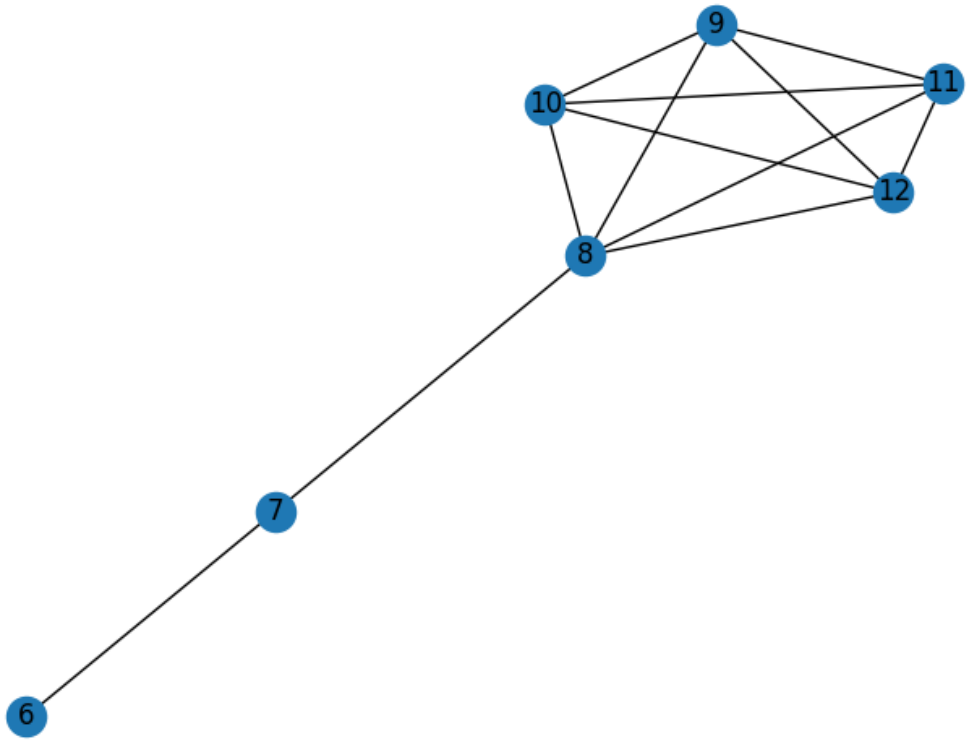

Figure 4.3: Second community

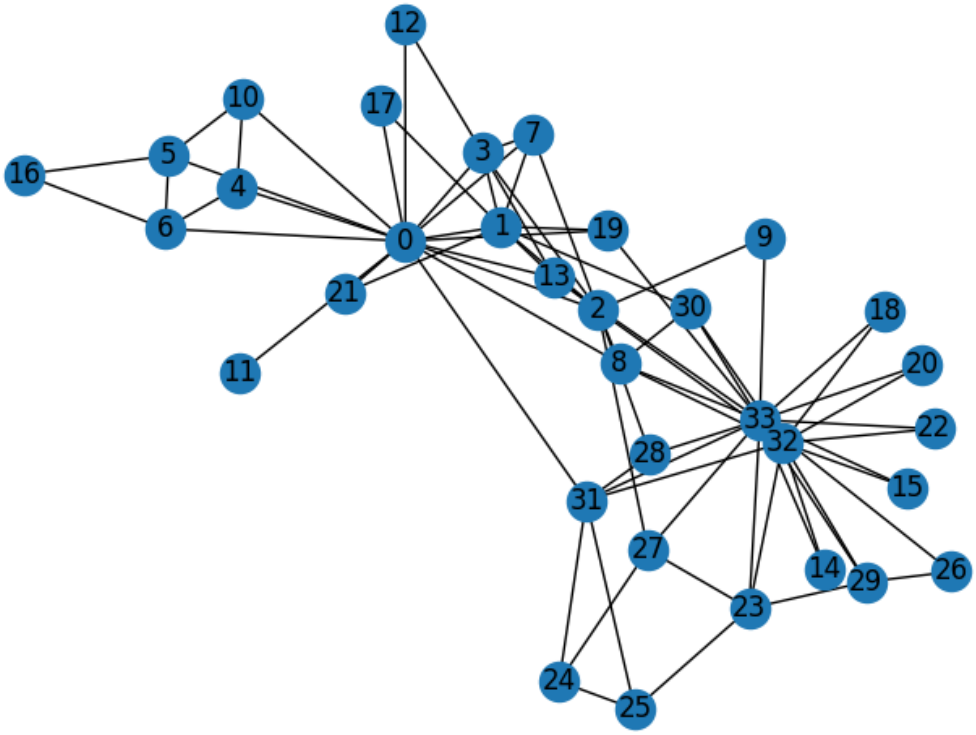

Figure 4.4: Zachary Karate club

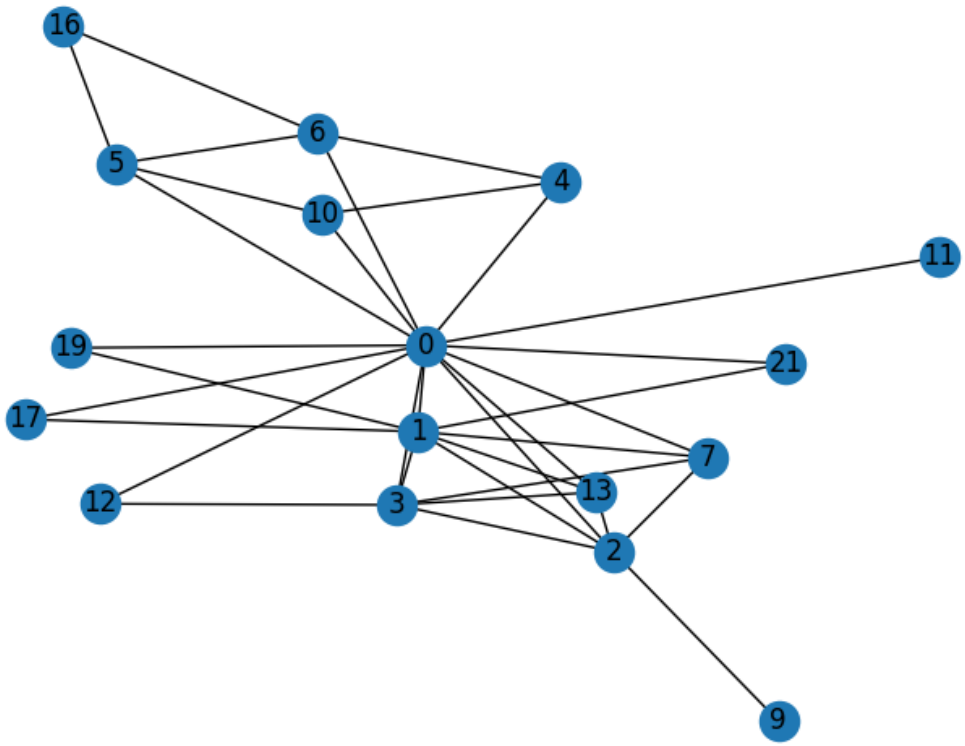

Figure 4.5:First community of Zachary Karate club

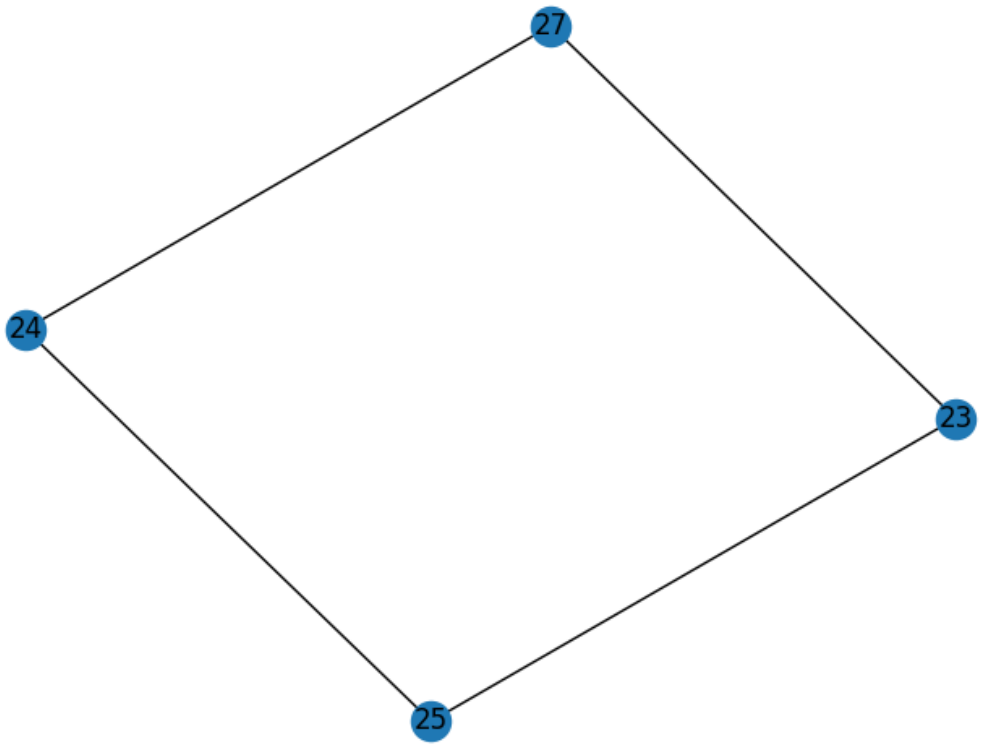

Figure 4.6: Second community of Zachary Karate club

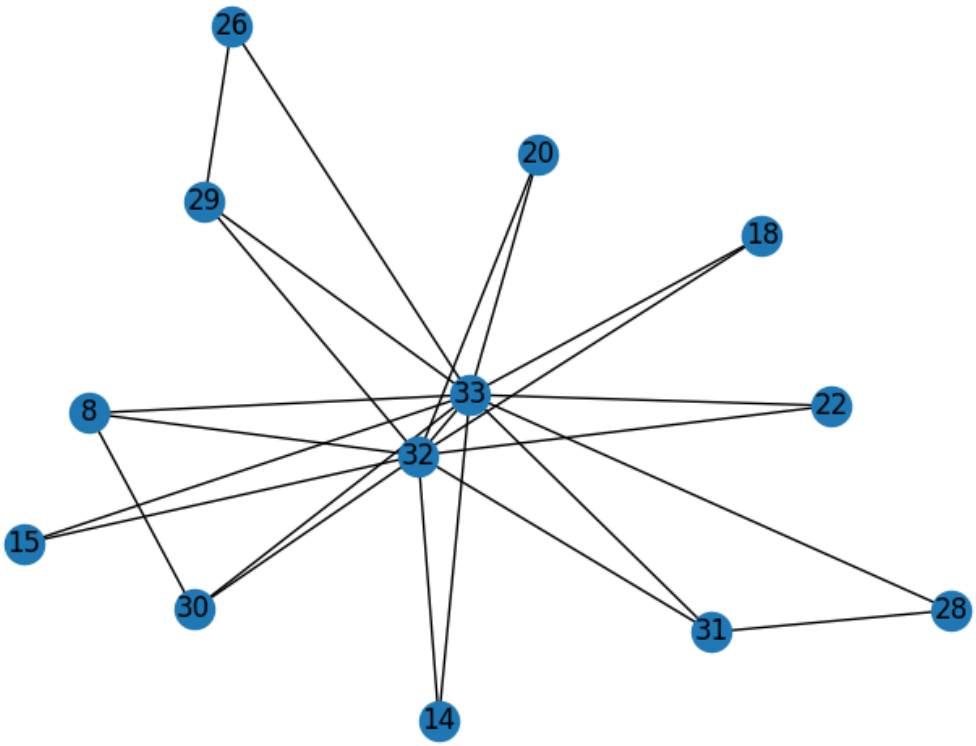

Figure 4.7: Third community of Zachary Karate club

Table 4.3: Modularity score of Fire propagation and boundary vertex model, Label propagation [35] and walked trap algorithms [36].

Algorithm	jazz Musicians	Zachary karate Club	Dolphin network	US political group	Relaxed Cavemen graph	LFR graph	Gaussian Random noise graph
FPB algorithm	0.40	**0.37253**	**0.518282**	**0.51**	0.6394949	**0.459493**	**0.3765**
Label prop	0.430390432	0.3532215	0.50007713	0.49962849	0.6453669	0.11456	0.364544
Walked Trapped	0.438421	0.3532215	0.48884537	0.50697240	0.64156049	0.3243	0.3567

As from the above result showed in the table we can easily found out that our proposed novel method is performed better than the currently presented method such as walked trapped method [35] and label propagation [36] the modularity score of our proposed method is better than walked trapped and label propagation algorithm on a real-world dataset such as US political book, Bottlenose dolphin, Zachary karate club datasets and also synthetic dataset such as LFR benchmark, Gaussian Random dataset. .

4.2.3 Computation complexity

Let the graph with $|V|$ number of vertex and $|E|$ edges, As our proposed method work in 2 stages in first stage our algorithm have find overlapped communities . Let $|C|$ be average number of communities developed in first stage so time complexity of first stage is O($|C|\times(|V| +|E|)$) then our method found all the boundary vertex that are belong to more than one community in O($|C|\times(|V|)$ time complexity and in in last stage we reassign the boundary vertex to their respective community using Dijkstra distance. Let $|B|$ be the total number of boundary vertex and $|C_v|$ be total number of core vertex then time complexity of reasinging the boundary vertex is O($|B|\times(|C_v|)$. So total time complexity of our proposed method is O($|C|\times(|V| +|E|)$) +O($|B|\times(|C_v|)$+O($|B|\times(|C_v|)$.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In the thesis, a new novel method for community detection is proposed named as fire propagation and boundary vertex algorithm based on a nature-inspired algorithm and boundary vertex concept which commences with selecting random vertices as the seed vertices and these vertices are used as the origin for fire fan out method this results in overlapping communities based on membership value calculated for each vertex then for the removal of overlapping communities boundary vertex concept is used which first find out the boundary vertices (vertices that belonging to more than one communities) then these vertices finally assigned to unique communities based on Dijkstra distance of each boundary vertices with seed vertices is calculated along with the count of adjacent vertices that belong to community originated for the particular seed vertices. Finally, the proposed method evaluated using modularity scores on the various dataset, and this modularity score is compared with walked trap and label propagation method modularity scores.

5.2 Future work

The proposed method (Fire propagation and boundary vertex) works well for an undirected graph and this work can be extended for directed graphs, weighted graphs, and weighted directed graphs. In terms of modularity also this work can be extended further to improve the partitioning of the graphs.

Appendix A

Appendices

A.1 List of publications (Communicated)

A.1.1 Community detection using Fire propagation and Boundary vertices

1Department of Computer Science and Engineering, Delhi Technological University,
Shahbad Daulatpur, Main Bawana Road
Delhi 110042, India
sanjay.kumar@dtu.ac.in, rahul1201dtu@gmail.com

Abstract: Community detection in complex networks deal with grouping related nodes together and plays a vital role to understand the functioning of the system in real-life situations. Community detection is classified as an NP-hard problem. Various algorithms are currently available for it but the problem with these existing algorithms is either they have high in time complexity or they have not able to partition the network perfectly. In this paper, we propose a novel community detection algorithm that works in two phases. In the first phase, we apply fire propagation technique in which choosing an arbitrary vertex as the core vertex and connecting an adjacent vertex to it and shapes a community this is similar to how fire spreads in real-life situations. In the second phase,we use the result of the first phase of an overlapped community and detect all boundary vertices which are belongings to more than one communities and assign them to the single community based on the weight that each core vertex assign to that particular boundary

vertex using Dijkstra distance and the count of the adjacent vertex that belong that community. The proposed algorithm performs well as compared to label propagation and walktrap algorithm in terms of modularity score using various synthetic and real-world datasets.

Keywords: Boundary vertices, complex network ,community detection ,core vertex ,fire propagation ,modularity score ,social networks

1. Introduction

A.1.2 Community detection algorithms in complex networks: A survey

1Department of Computer Science and Engineering, Delhi Technological University,
Shahbad Daulatpur, Main Bawana Road
Delhi 110042, India
sanjay.kumar@dtu.ac.in, rahul1201dtu@gmail.com

Abstract: Community detection in the complex network is the process of finding optimal clusters of vertices that are similar in characteristics. To study the properties and functions of complex networks, community detection plays a crucial role. Community detection is generally categorized as an optimization problem, and due to its inherent property it can't be solved by the traditional optimization method, and over the past few decades, various algorithms have been proposed to address this problem in multiple fields such as power system, physics, biology, or sociology. In this paper, we present a critical survey on various algorithms for community detection currently available such as genetic algorithm, evolutionary algorithms, a nature-inspired algorithm, deep learning algorithm. This survey paper outlines the challenges and constraints of different state-of-arts community algorithms detection by utilizing contemporary techniques like deep neural networks, genetic algorithms, and various topological features based methodologies.

Keywords: Community detection, Complex network, Deep learning algorithm, Genetic algorithm, Girvan Newman algorithm, Modularity score

Bibliography

[1] Karataş, Arzum, and Serap Şahin. "Application areas of community detection: A review." 2018 International congress on big data, deep learning and fighting cyber terrorism (IBIGDELFT). IEEE, 2018.

[2] J. C. Valverde-Rebaza, and A. de Andrade Lopes, "Link prediction in complex networks based on cluster information," Proc. 21 st Brazilian Conf. in Artificial Intelligence,2012, pp. 92-101.

[3] S. Soundarajan, and J. Hopcroft, "Using community information to improve the precision of link prediction methods," in Proc. 21st Int.Conf. on World Wide Web,2012, pp. 607-608.

[4] H. Sarvari, E. Abozinadah, A. Mbaziira, and D. Mccoy,"Constructing and analyzing criminal networks," in 2014 IEEE Security and Privacy Workshops , 2014, pp. 84-91.

[5] A. Karataş, and S. Şahin, "A Review on Social Bot Detection Techniques and Research Directions," in Proc. Int. Security and Cryptology Conference Turkey, 2017, pp.156-161.

[6] M. Salathé, and J. H. Jones, "Dynamics and control of diseases in networks with community structure," PLoS Computational Biology, vol. 6, no. 4, pp. e1000736, 2010.

[7] J. J. Bechtel, W. A. Kelley, T. A. Coons, M. G. Klein, D. D. Slagel,and T. L. Petty, "Lung cancer detection in patients with airflow obstruction identified in a primary care outpatient practice," Chest, vol. 127, no. 4, pp. 1140-1145, 2005.

[8] N. Haq, and Z. J. Wang, "Community detection from genomic datasets across human cancers" 2016 IEEE Global Conf. on Signal and Infor. Process., 2016,pp. 1147-1150.

[9] Y. Yang, P. G. Sun, X. Hu, and Z. J. Li, "Closed walks for community detection," Physica A: Statistical Mechanics and its Applications, vol. 397, pp. 129-143, 2014.

[10] A. Karataş, and S. Şahin, "A Review on Social Bot Detection Techniques and Research Directions," in Proc. Int. Security and Cryptology Conference Turkey, 2017, pp.156-161.

[11] A. Bienkov, "Astroturfing: what is it and why does it matter?,"[Online]. Available:https://www.theguardian.com/commentisfree/2012/feb/08/what-is-astroturfing

[12] M. J. Mosadegh, and M. Behboudi, "Using social network paradigm for developing a conceptual framework in CRM," Australian J. Bus.and Manage. Research, vol. 1, no. 4, pp. 63, 2011.

[13] C. Remy, B. Rym, and L. Matthieu, "Tracking bitcoin users activity using community detection on a network of weak signals," in Int. Workshop on Complex Networks and their Applications, 2017,pp. 166-177.

[14] N. Kokkiligadda, and V. K. Vatsavayi, "Community privacy preservation in dynamic social networks.", Int. J. Tech. Research and Applicat., vol. 4, no.6 , pp. 133-136, 2016.

[15] Campan, Y. Alufaisan, and T. M. Truta, "Community Detection in Anonymized Social Networks," in Proc. Workshops of the EDBT/ICDT 2014 Joint Conference, 2014, pp. 396-405.

[16] S. B. Abdrabbah, R. Ayachi, and N. B. Amor, "Collaborative filtering based on dynamic community detection," Dynamic Networks and Knowledge Discovery, vol. 85, 2014

[17] D. Lalwani, D. V. Somayajulu, and P. R. Krishna, "A community driven social recommendation system," in Proc. 2015 IEEE Int.Conf. on Big Data, 2015, pp. 821-826.

[18] P. Bródka, P. Kazienko, and B. Kołoszczyk, "Predicting group evolution in the social network," in Int. Conf. on Social Informatics, 2012, pp. 54-67.

[19] G. Diakidis, D. Karna, D. Fasarakis-Hilliard, D. Vogiatzis, and G. Paliouras, "Predicting the evolution of communities in social networks," in Proc. 5th Int. Conf. on Web Intelligence, Mining and Semantics, 2015, p. 1.

[20] Guerrero, Manuel, et al. "Adaptive community detection in complex networks using genetic algorithms." Neurocomputing 266 (2017): 101-113.

[21] Chapela V et al (2015) Mathematical foundations: complex networks and graphs (a review). Intentional risk management through complex networks analysis. Springer, Cham, pp 9–36

[22] Brandes U (2008) On variants of shortest-path betweenness centrality and their generic computation. Soc Netw 30(2):136–145

[23] Jaccard P (1901) ETude comparative de la distribution florale dans une portion des Alpes et des Jura. Bull Soc Vaudoise Sci Nat 37:547–579

[24] Guo, Kun, et al. "A local community detection algorithm based on internal force between nodes." Applied Intelligence 50.2 (2020): 328-340.

[25] Holland, John Henry. Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. MIT press, 1992. U. N. Raghavan, R. Albert, and S. Kumara,Physical Review E 76, 036106 (2007)

[26] Pattanayak, Himansu Sekhar, Amrit Lal Sangal, and Harsh K. Verma. "Community detection in social networks based on fire propagation." Swarm and evolutionary computation 44 (2019): 31-48.

[27] Gui, Qiong, et al. "A community discovery algorithm based on boundary nodes and label propagation." Pattern Recognition Letters 109 (2018): 103-109.

[28] P.M. Gleiser, L. Danon, Community structure in jazz, Adv. Complex Syst. 6 (2003) 565–573.

[29] W.W. Zachary, An information flow model for conflict and fission in small groups, J. Anthropol.Res. 33 (1977) 452–473.

[30] 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 (2003) 396–405.

[31] V. Krebs, Orgnet, 2017, http://www.orgnet.com.

[32] A. Lancichinetti, S. Fortunato, Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities, Phys. Rev. E 80 (2009) 016118.

[33] D.J. Watts, Small Worlds: the Dynamics of Networks between Order and Randomness, Princeton University Press, 1999.

[34] M.E. Newman, M. Girvan, Finding and evaluating community structure in networks, Phys. Rev.E 69 (2004) 026113.

[35] P. Pons, M. Latapy, Computing communities in large networks using random walks, J. Graph Algorithm Appl. 10 (2006) 191–218

[36] U. N. Raghavan, R. Albert, and S. Kumara,Physical Review E 76, 036106 (2007)

[37] T. Falkowski, J. Bartelheimer, M. Spiliopoulou, Mining and visualizing the evolution of subgroups in social networks, in: IEEE/WIC/ACM International Conference on Web Intelligence, 2007, pp. 52–58.

[38] X. Ma, D. Dong, Evolutionary nonnegative matrix factorization algorithms for community detection in dynamic networks, IEEE Trans. Knowl. Data Eng. 29(5) (2017) 1045–1058.