International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-1, October, 2019

Community Detection Algorithm in Social Networks through Iterative Analysis based on Degree of Nodes

Amedapu Srinivas, R. Leela Velusamy

Abstract— Social networking is the grouping of individuals into specific groups, like small rural communities or a neighborhood subdivision. A fundamental problem in the analysis of social networks is the tracking of communities. A community is often defined as a group of network members with stronger ties to members within the group than to members outside the group. The traditional method for identifying communities in networks is hierarchical clustering. Recently, several works have been done in this community identification using different type of clustering algorithm and connectivity-based scoring function. In this paper Random Head Node Technique and Highest Degree Head Node Techniques are proposed to group the nodes into communities. In these techniques best set of centroids are chosen based on the fitness value to cluster the nodes into communities.

Index Terms— network node, clustering, random head node, highest degree head node, social networks, social network community, cluster, community detection.

I. INTRODUCTION

A social network (SN) for a single person is made with her/his alliance and individual associations with different individuals in the general public. SNs speak to and model the social bind among people. With the quick development of the web, there is a huge development in online communication of the clients. Numerous person to person communication destinations, e.g., Facebook, Twitter and so forth have additionally come up to encourage client alliance. As the quantity of alliances have expanded complex, it is getting to be hard to monitor these correspondences. Individuals will in general get related with individuals of comparable likings and tastes. The simple to-utilize web-based social networking enables individuals to expand their public activity in extraordinary ways since it is hard to meet companions in the physical world, however a lot simpler to discover companions online with comparable interests. These true SNs have intriguing examples and properties which might be investigated for various valuable purposes.

A1021109119

SNs have a trademark property to display a network structure. In the event that the vertices of the system can be divided into either covering or disjoint sets of vertices to such an extent that the quantity of links inside a group surpasses the quantity of links between any 2 sets by sensible sum, we state that the system shows a connecting structure. Systems showing a network structure may frequently display a various leveled network structure also.

Thee way toward finding the strong groupings or bunches in the system expressed as community detection (CD). It is one of the key errands of SN examination. The recognition of networks in SNs can be helpful in numerous areas where cooperative choices are taken, e.g. multicasting an information important to a network as opposed to sending it to everyone in the grouping or prescribing a lot of items to a network.

In this paper, Random Head Node Technique and Highest Degree Node Head Technique (HDNHT) are proposed to cluster the similar community nodes. In Random Head Node Technique every time n centroid nodes are taken randomly and the nodes are clustered based centroid taken and the fitness is calculated. In Highest Degree Node Head Technique (HDNHT) technique, the n centroid nodes are taken randomly in the first iteration and the nodes are clustered based on the centroid taken and the fitness is calculated. From the next iteration, the centroid nodes are selected based on the degree of the node. After updating new centroids, the clustering process is done and the fitness is calculated based on the connections and the same process is repeated as many times as required. After all iterations based on the fitness value best centroids will be chosen.

II. RELATED WORK

2.1 Fundamental Concepts

2.1.1 Social Network (SN)

A SN is depicted by SN chart G containing n number of hubs showing n individuals or the individuals in the framework. The relationship between hub i and hub j is addressed by the edge e_{ij} of the graph. A coordinated or an undirected outline may speak to these relationships between the individuals from the framework. The diagram can be addressed by a continuity system A wherein Aij = 1 in case there is an edge among I and j else Aij = 0.

Revised Manuscript Received on October 30, 2019.

* Correspondence Author

Amedapu Srinivas, Full Time Ph.D. Research Scholar, Department of Computer Science and Engineering, National Institute of Technology, Tiruchirappalli, Tamil Nadu, INDIA – 620015. Email: SrinivasReddyAmedapu@gmail.com

R. Leela Velusamy, Professor, Department of Computer Science and Engineering, National Institute of Technology, Tiruchirappalli, Tamil Nadu, INDIA – 620015. Email: leela@nitt.edu

Some examples [1] of SNs incorporate companions based, phone, email and joint effort systems. These systems can be spoken to as charts and it is attainable to ponder and investigate them to discover intriguing examples among the substances. These engaging models can be used in different helpful applications.

2.1.2 Community

A community can be characterized as a grouping of elements closer to one another in contrast with different elements of the dataset. System is made by individuals to such a degree, that those inside a system participate with each other more as regularly as conceivable as with those outside the system. The closeness between substances of a grouping can be estimated by means of likeness or separation measures between elements. McPherson et al [2] expressed that "closeness breeds association". They examined different social components which lead to comparable conduct or homophily in systems. The people group in SNs are closely resembling bunches in systems. An individual spoken to by a node in charts may not be a piece of only a network or a grouping, it might be a component of numerous intently related or various groupings existing in the system. For instance an individual may simultaneously have a place with school, school, loved ones groupings. Every single such network which have basic nodes are called covering networks.

Verification and assessment of the system structure has been done by various specialists applying frameworks from different kind of examinations. The idea of collection in frameworks is normally settled on a choice by gathering coefficient which is an extent of how much the vertices of a framework will all in all gathering at one spot. The neighborhood bunching coefficient [3] and worldwide grouping coefficient [4] are two sorts of collection coefficients discussed recorded as a hard copy.

2.1.3 A technique: clustering similar items

Networks are those pieces of the diagram which have denser associations inside and couple of associations with the remainder of the graph8. The aim of unsupervised learning is to assemble comparable articles with no earlier information about them. If there should be an occurrence of systems, the bunching issue alludes to grouping of nodes as per their closeness registered dependent on topological highlights or potentially different attributes of the diagram. System parceling and bunching are two normally utilized techniques in writing to discover the groupings in the SN chart. These techniques are quickly depicted in the coming sections.

A. Diagram dividing

Diagram dividing is the path toward distributing a chart into a predefined number of smaller parts with unequivocal properties. An average property to be restricted is called cut size. A cut is a package of the vertex set of a chart into two disjoint subsets and the component of the cut is the amount of edges between the portions. A multicut is a great deal of edges whose departure isolates the chart into at any rate two sections. It is imperative to demonstrate the amount of sections one wishes to get if there ought to emerge an event of diagram apportioning. The size of the sections ought to in like manner be shown, as for the most part a plausible yet not significant course of action is put the base degree vertex into one section and the rest of the vertices into other. Since the amount of systems is ordinarily not known early, diagram dividing methods are not sensible to recognize organizes in such cases.

B. Clustering

Clustering is the way toward grouping a lot of comparable things in structures known as clusters. Clustering the SN diagram may give a great deal of data about the basic concealed traits, connections and properties of the members just as the associations among them. Various leveled grouping and dividing strategy for clustering are the regularly utilized clustering systems utilized in writing.

In hierarchial clustering, a chain of command of bunches is framed. The procedure of chain of command creation or leveling can be agglomerative or disruptive. In agglomerative bunching strategies, a base up way to deal with grouping is pursued. A specific node is clubbed or agglomerated with comparative nodes to shape a group or a network. This total depends on comparability. In diverse clustering approaches, a huge bunch is over and over isolated into littler groups.

Methods of partition start with an initial partition in between the number of clusters pre-set and relocation of specimens by moving them across clusters, e.g., K-means clustering. A comprehensive assessment of every conceivable segment is required to accomplish worldwide optimization in clustering based on partition. This is tedious and here and there infeasible, thus scientists utilize insatiable heuristics for iterative enhancement in dividing techniques for clustering. The following segment sorts and talks about real calculations for CD.

2.1.4 Algorithms for Community Detection

Various people group location calculations and techniques have been proposed and conveyed for the distinguishing proof of networks in writing. There have likewise been alterations and updates to numerous techniques and calculations previously proposed. A total investigation of CD in outlines has been done by Fortunato[5] in the year 2010. Various reviews available recorded as a hard copy are by Coscia et al [6] in 2011, Fortunato et al [7] in 2012, Porter et al [8] in 2009, Danon et al [9] in 2005, and Plantié et al [10] in 2013.The exhibited work audits the calculations accessible till 2015 as far as we could possibly know incorporating the calculations given in the before overviews. Papers dependent on new methodologies and systems like enormous information, not examined by past writers have been consolidated in our article.

A. Community Detection: Graph Partitioning

Graph partitioning based strategies have been utilized in literature to separate the graph into parts with the end goal that there are not many associations between segments. The Kernighan-Line [11] algorithm for *Graph partitioning* was among the soonest procedures to partition a diagram or graph. It partitions the nodes of the diagram with expense on edges into subsets of given sizes in order to limit the total of expenses on all edges cut. A noteworthy impediment of this calculation anyway is that the quantity of groupings must be predefined.

The calculation anyway is very quick with a most pessimistic scenario running time of ($n2$). Newman [12] diminishes the broadly read greatest probability strategy for network identification to an inquiry through a grouping of up-and-comer arrangements, every one of which is itself an answer for a base cut diagram dividing issue.

B. Community Detection: Clustering Based

The fundamental worry of network location is to recognize bunches, groupings or firm subgroups. The premise of an enormous number of network identification calculations is grouping. Among the pioneers of network identification strategies, Girvan and Newman [13] had a principle job. They proposed an inconvenient estimation subordinate on edge betweenness for a graph with undirected and unweighted joins. The computation focused on edges that are most "between" the systems and systems are grown persistently by ousting these connections from the main outline. Three unique measures for count of edge-betweenness in vertices of a diagram were proposed in Girvan and Newman [14].

Table-1. Community Detection Based on Clustering

S.No.	Approach
1	Girvan and newman[14] proposed an approach called Divisive clustering which contains the parameters{edge betweenness}
2	Newman[15] [16] [17] proposed an approach of modularity maximization which contains theparameters as{Modularity, Eigenvector and Eigen value }
3	Clauset e al [18] proposed Greedy optimization of modularity. It contains the parameters as{ vertices , Edges, Modularity }
4	Blondel et al (Louvain Method) [19] proposed an approach of Hierarchical clustering. It contains the parameters as{ Modularity , Nodes, edges }
5	Zhou et al [21], Guimera et al [20], proposed optimization of Modularity. It contains the parameters as{37: No. of links,: No. of edges, inter factor and intra factor, Modularity, no. of partitions, Modularity, linking probability, no. of modules}
6	Duch et al [22] proposed optimization of Modularity. It contains the parameters as{ degree, Modularity , links , No. of nodes, }
7	Ye et al (AdClust) [23] proposed an approach called Agglomerative Clustering. It contains the parameters as{ Force, Modularity , Vertices, }
8	Sheppard [24] and Wahl proposed an approach called Hierarchical Fuzzy Spectral clustering .It contains the parameters as{ Jaccard Similarity , Fuzzy modularity }
9	Falkowski et al (DENGRAPH) [25] proposed an approach called clustering based on density. It contains the parameter named as Distance Function.
10	Dongen et al (MCL) [26] proposed an approach called Markovian Clustering. . It contains the parameter named as number of nodes.

The most pessimistic scenario time intricacy of the link betweenness calculation is ($m2n$) and is ($n3$) for meager charts, where m means the quantity of edges and n is the quantity of vertices.

The improvement of measured quality capacity has gotten incredible consideration in writing. The Table .1 records bunching based network location strategies, including algorithms which use measured quality and seclusion advancement.

C. Community Detection: Genetic Algorithms (GA)

Genetic algorithms (GA) are versatile heuristic hunt calculations whose point is to locate the best arrangement under the given conditions. A hereditary calculation begins with a lot of arrangements known as chromosomes and fitness function is determined for these chromosomes. On the off chance that an answer with a most extreme fitness is gotten, one stops else with some likelihood hybrid and transformation administrators are connected to the present arrangement of answers for acquire the new arrangement of arrangements. Network discovery can be seen as an improvement issue in which a target work that catches the instinct of a network with preferred inside availability over outside network is picked to be upgraded. GA has been connected to the procedure of network disclosure and investigation in a couple of ongoing examination works. These are portrayed quickly in this segment. Table .2 enrolls the calculations accessible in writing for network location dependent on GA.

Table-2. Community Detection using Genetic Algorithms

S.No.	Approach
1	Pizzuti(GA-Net) [27] proposed an approach called Community score as fitness function which contains the parameters{ Community score }
2	Pizzuti(MOGA-Net)[28] proposed an approach called optimization using multi objective which contains the parameters{ Community fitness, Community score, }
3	Hafez et al [29] proposed an approach called multi objective optimization, Single objective which contains the parameters{ mutation Crossover operators , Number of genes }
4	Mazur et al [30] proposed an approach called Community score and Modularity as fitness functions using the parameters of Fitness functions
5	Liu et al [31] proposed an approach called Genetic algorithm and Clustering which contains the parameters{ maximal generation number, Size of population, greatest no. of ages for unchanged fittest chromosome part of mined hubs , no. of networks, }
6	Tasgin et al [32] proposed an approach called Clustering and Genetic algorithm which contains the parameters{ number of chromosomes , Modularity , population size }
7	Zadeh [33] proposed an approach called Multi population cultural algorithm which contains the parameters{ BSN , BS_average }

D. Community Detection: Label Propagation

Name or number engendering in a system is the spread of a number to different centers existing in the system. Every center point accomplishes the number controlled by a greatest number of the neighboring hubs. This area talks about some name engendering based strategies for finding networks.

Table-3. Tabular Algorithms list

S.No.	Approach
1	Raghavan et al(LPA) [34] proposed an approach called propagation with iterative label which contains the parameters{ 55: nodes, labels 56: labels, threshold 57: label, similarity 59: nodes }

2	Xie et al(LabelRank) [35] proposed an approach called Inflation cut- off Conditional update, propagation which contains the parameters{ belongingness coefficient, threshold, nodes }
3	Wu et al(BMLPA) [36] proposed an approach called Overlapping Communities, Label propagation which contains the parameters{ labels to which vertices belong, average degree, Number of vertices }

E. Community Detection: Semantics

Semantic content and edge connections in a semantic network might be furthermore used to segment the nodes into networks. The context, just as the relationship of the nodes, both are thought about during the time spent semantic CD. LDA(Latent Dirichlet Allocation)[37] is utilized in a few semantic network based network identification approaches. A grouping calculation dependent on the Link-field-Topic (LFT) model is advanced by Xin et al [38] to conquer the confinement of characterizing the quantity of networks already present. The examination shapes the semantic link weight (SLW) in light of the examination of LFT, to assess the semantic load of connections for each inspecting field. The proposed clustering algorithm depends on the SLW which could isolate the semantic SN into bunching units. In another work [39] the creators have utilized ARTs model and separated the procedure into two stages to be specific LDA sampling and network identification. In the past method diverse testing ARTs have been arranged. A people group bunching calculation has furthermore been proposed. The system could recognize the present covering systems. Xia et al [40] built up a semantic framework using information from the comment substance isolated from the basic HTML source records. An ordinary score is gained for two customers for every association tolerating comments to be undeniable associations between people. An interpretive technique for taking out comment substance is proposed to build the semantic framework for example, the terms and articulations in data are incorporated into comments as solid or repudiating.

III. EPROPOSEDE COMMUNITY GROUPING TECHNIQUES

This section explains the proposed community grouping methods called Highest Degree Node Head Techniques (HDNHT) as well as Random Head Technique.

Random Head Technique – In this technique n number of nodes is chosen randomly as centroids from the network or graph. Thereafter the shortest path between each centroid and each node will be calculated and then the nodes will be grouped into a cluster based on the shortest distance to the centroid. And then the fitness has to be calculated for each cluster and summation of fitness of all clusters will be noted. This process will be iterated for chosen number of times, with the selection of other set of n centroids every time. After the final iteration, the set of centroids those obtained the best fitness value is chosen to cluster the nodes into communities.

Highest Degree Node Head Technique – In this technique, initially, like in Random-Head Technique, n number of nodes is chosen randomly as centroids from the network or graph. And nodes will be clustered based on shortest distance to the centroids, and then fitness value will be calculated. Thereafter, the new n nodes, one node from each cluster based on the highest degree are chosen as centroids for the next iteration. This process will be iterated for chosen number of times. After the final iteration, the set of centroids those obtained the best fitness value is chosen to cluster the nodes into communities

A. Random Head Technique

In Random Head Technique, initially n number of centroids is randomly selected from the graph/network. Consider the Graph shown in Fig-1 for the discussion of Random Head Technique:

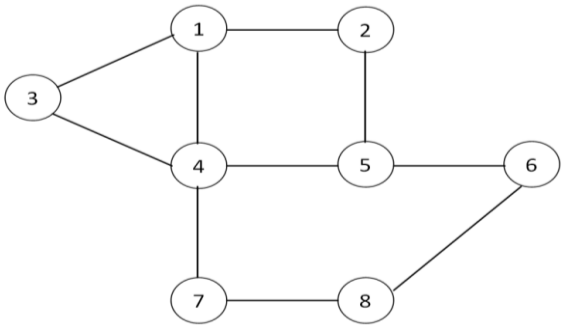

Fig-1: Sample Connected Graph

Let us assume, initial random selected centroids are node-1 and node-6 as centroid-1 and centroid-2 respectively, then the shortest path between these centroids to all the nodes in the graph has to be calculated to make communities. The following Table-4 shows the shortest path from each centroid to all other nodes in the graph.

After calculating the distance from each centroid to all other nodes in the graph, group the nodes to form communities based on the lowest distance between the centroids and the nodes. For example, the distance between the node-1 and centroid-1 is 0 and the distance between the node-1 and centroid-2 is 3, so the node-1 will be associated with centroid-1, as well as the distance between node-2 and centroid-1 is 1 and the distance between node-2 and centroid-2 is 2, means the node-2 will be associated with centroid-1 and the distance between node-5 and centroid-1 is 2 and node-5 and centroid-2 is 1, so node-5 will be associated with centroid-2. When a node is at same distance from more than one centroid, it can be associated with any of the centroids. Example node-7 is at a distance of 2 from centroid-1 as well as centroid-2, so it can be associated with either cetriod-1 or centroid-2. Table-5 shows the centroids and associated nodes with them, simply called as communities.

Table-4. Distance from centroids

Centroid / Node No.	Centroid-1 (node-1)	Centroid-2 (node-6)
1	0	3
2	1	2
3	1	3
4	1	2

5	2	1
6	3	0
7	2	2
8	4	1

Now the total fitness (explained in section 3.2) of these communities will be calculated before going to the next iteration.

Table-5. Clusters of centroids

Centroid-1 (node-1)	Centroid-2 (node-6)
1	5
2	6
3	8
4	
7	

For the next iteration the randomly one node is nominated from each community. This process will be repeated as many times as we set or will be stopped when same set of nodes are nominated as centroids by the algorithm.

B. Highest Degree Node Head Technique (HDNHT)

In this Highest Degree Node Head Technique, initially n number of centroids is randomly selected from the graph/network. Consider the Graph shown in Fig-1 for the discussion of HDNHT technique:

Let us assume, initial random selected centroids are node-2 and node-7 as centroid-1 and centroid-2 respectively, then the shortest path between these centroids to all the nodes in the graph has to be calculated to make communities. The following Table-6 shows the shortest path from each centroid to all other nodes in the graph.

After calculating the distance from each centroid to all other nodes in the graph, group the nodes to form communities based on the lowest distance between the centroids and the nodes. For example, the distance between the node-1 and centroid-1 is 1 and the distance between the node-1 and centroid-2 is 2, so the node-1 will be associated with centroid-1, as well as the distance between node-8 and centroid-1 is 3 and the distance between node-8 and centroid-2 is 1, means the node-8 will be associated with centroid-2. When a node is at same distance from more than one centroid, it can be associated with any centroid. Table-7 shows the centroids and associated nodes with them, simply called as communities.

Table-6. Distance from centroids

Centroid / Node No.	Centroid-1 (node-2)	Centroid-2 (node-7)
1	1	2
2	0	3
3	2	2
4	2	1
5	1	2
6	2	2
7	3	0
8	3	1

Now the total fitness (explained in section 3.2) of these communities will be calculated before going to the next iteration. For the next iteration the highest degree node from each community will be nominated as centroid, i.e., as per the considered example either node-1 or node-5 by the community-1 and node-4 by the community-2 are nominated as centroids for the next iteration. This process will be repeated as many times as we set or will be stopped when same set of nodes are nominated as centroids by the algorithm.

Table-7. Clusters of centroids

Centroid-1 (node-2)	Centroid-2 (node-7)
1	4
2	7
3	8
5	
6	

C. Fitness Calculation

The fitness is calculated based on the connection between the node to the other nodes, whether they are within the community or outside/other community. The following equation (1) is used to calculate the fitness node and community:

$$fit_k = \sum_{i=1}^{I}\left[\frac{1}{J}\sum_{j=1}^{J} no.\ of\ connection\ \ b/w\ \ j\ \ and\ \ nodes\ \ outside\ \ i\right] \rightarrow (1)$$

Where,

fit_k → Fitness of centroids in k^{th} iteration

I → Total number of clusters

J → Total number of nodes in i^{th} cluster y,

The *no. of connection b/w j and nodes outside i* in the above equation represents the number of connections between j^{th} node in i^{th} cluster and other nodes in the clusters outside i^{th} cluster. The fitness calculation is explained as follows: In the Table-2 we have two clusters, the nodes in the first cluster are 1, 2, 3, 5 and 6; and the nodes in the second cluster are 4, 7 and 8. Here while considering the first cluster, the connections between node-1 and node-2, node-1 and node-3, and node-1 and node-4 have to be checked. If there is connection to the other cluster member put the value as one; otherwise zero.

Here, the connection between node-1 and node-2, node-1 and node-3 are within the same cluster, so the value is 0, where as the connection between the node-1 and node-4 is outside the node-1's cluster, therefore the value is 1, i.e.,

$$C_1(n1) = 0 + 0 + 1 = 1$$

In the above equation, $C_1(n1)$ denotes the connection between node-1 of C_1 to the other cluster nodes. Similarly the following equations show the other nodes connections of cluster-1.

$$C_1(n2) = 0 + 0 = 0$$
$$C_1(n3) = 0 + 1 = 1$$
$$C_1(n5) = 0 + 1 + 0 = 1$$
$$C_1(n6) = 0 + 1 = 1$$

Thereafter, the value of C_1 has to be calculated which is the ratio of sum of the obtained values for all nodes in the cluster to the total number of nodes in the cluster. It is shown by an equation below:

$$C_1 = \frac{C_1(n1) + C_1(n2) + C_1(n3) + C_1(n5) + C_1(n6)}{5}$$

$$C_1 = \frac{1 + 0 + 1 + 1 + 1}{5} = \frac{4}{5}$$

Similarly, the value for C_2 has to be calculated using its respective nodes.

$$C_2(n4) = 1 + 1 + 0 + 1 = 3$$
$$C_2(n7) = 0 + 0 = 0$$
$$C_2(n8) = 1 + 0 = 1$$
$$C_2 = \frac{C_2(n4) + C_2(n7) + C_2(n8)}{3}$$
$$C_2 = \frac{3 + 0 + 1}{3} = \frac{4}{3}$$

The fitness is then calculated by summing the values of C_1 and C_2 which is given by an equation as follows:

$$fit_k = C_1 + C_2$$

$$fit_1 = \frac{4}{5} + \frac{4}{3} = \frac{32}{15} = 2.133$$

The less value denotes better fitness and the selection of best fitness is shown by an equation below:

$$best(fit) = \min\{(fit_1), (fit_2), (fit_3), (fit_4), \ldots (fit_k)\} \rightarrow (2)$$

Input: Network Nodes
Output: Clustered Similar Nodes

1. ***Start***
2. Select n number of nodes as centroids randomly
3. *For* each centroid
4. *For* each node
5. Calculate shortest path
6. *End for*
7. *End for*
8. Cluster based on shortest path
9. *For* each cluster
10. Calculate fitness
11. *End for*
12. *For* each node
13. Check the number of connections
14. *End for*
15. Choose a node with maximum number of connections from each cluster as a centroid for next iteration
16. Repeat step 3 to step 15 until final iteration
17. The best set of centroid is chosen based on best fitness
18. Similar nodes are clustered based on best set of centroids
19. ***Stop***

Fig-2: Algorithm - Highest Degree Node Head technique

IV. EXPERMENTAL RESULTS AND DISCUSSION

This section explains the results obtained for the techniques discussed in this paper. The obtained results from Random Node Head and Highest Degree Node Head are compared in terms of fitness, separability, density, execution time and memory taken for execution. The experimentation is implemented in java (jdk 1.7) that runs on the system that has the configuration as follows: windows 7, 32bit, core 2 duo with clock speed of 2.94GHz and has 2GB of RAM.

A. Dataset Description

The dataset used for our experimentation is taken from (http://snap.stanford.edu/data/wiki-Vote.html) "Wikipedia Vote Network". Wikipedia is a free encyclopedia written collaboratively by volunteers around the world. A small part of Wikipedia contributors are administrators, who are users with access to additional technical features that aid in maintenance. In order for a user to become an administrator a Request for adminship (RfA) is issued and the Wikipedia community via a public discussion or a vote decides who to promote to adminship. Using the latest complete dump of Wikipedia page edit history (from January 3 2008) they extracted all administrator elections and vote history data. This gave us 2,794 elections with 103,663 total votes and 7,066 users participating in the elections (either casting a vote or being voted on). Out of these 1,235 elections resulted in a successful promotion, while 1,559 elections did not result in the promotion. About half of the votes in the dataset are by existing admins, while the other half comes from ordinary Wikipedia users. The network contains all the Wikipedia voting data from the inception of Wikipedia till January 2008.

B. Evaluation Metrics

Separability – The separability denotes that good communities are well separated from the rest of the network. The separability is calculated by the following equation-3, which is the ratio of number of connections between the nodes in the i^{th} cluster itself, to the number of connections between the nodes in the i^{th} cluster and the nodes outside i^{th} cluster. The i denotes the total number of clusters.

$$Separability = \sum_{i=1}^{I} \frac{no.\ of\ inner\ connection}{no.\ of\ outer\ connection} \rightarrow (3)$$

Density – The density metric indicates that the good communities are well connected. The density is calculated by using equation-4, whereas CS represents cluster size.

$$Density = \sum_{i=1}^{I} \frac{no.\ of\ inner\ connection}{CS \times (CS - 1)} \rightarrow (4)$$

The execution time is the total time taken by each technique to execute the process; and the memory consumption indicates total memory taken by each technique to execute the process and the fitness calculation is explained in section 3.2.

C. Performance Comparision

Fitness comparison is shown in Fig-3, when number of iterations is five the fitness values are 4.29 and 3.10 for Random and HDNHT techniques respectively and when set to ten iterations the values are 4.11 and 3.60, finally when set to twenty five iterations the values are 3.42 and 2.86. These values show that the HDNHT technique performs better than Random technique as lower fitness is good for community.

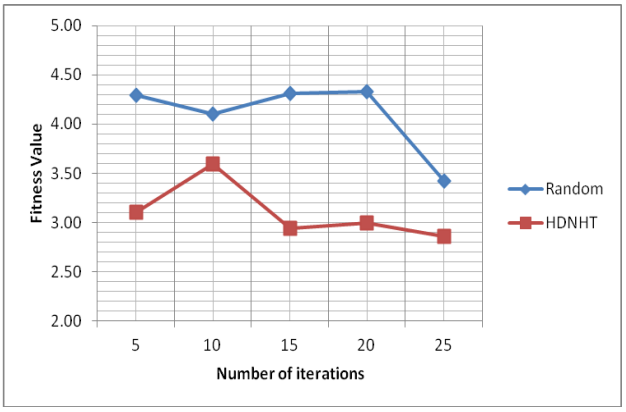

Fig-3: Fitness Comparison

The Fig-4 shows the separability comparison among Random node head technique and proposed technique HDNHT. Here when number of iterations are set to five the separability values are 0.94 and 1.23 for Random and HDNHT techniques respectively and when set to fifteen the values are 0.95 and 1.47, finally when set to twenty five iterations the values are 1.30 and 1.50. These values shows that HDNHT technique performs better than Random technique as higher separability value indicate that communities are well separated.

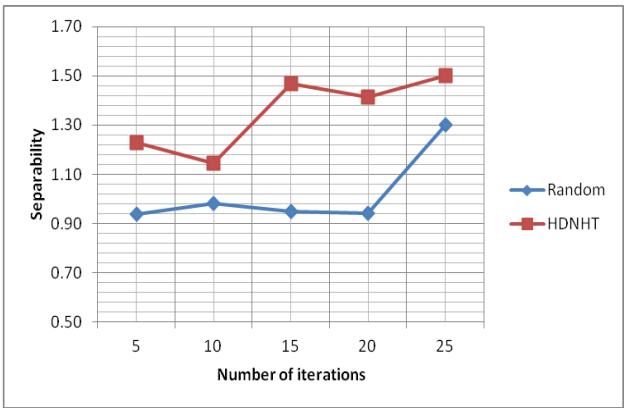

Fig-4: Separability Comparison

The Fig-5 shows the density comparison among the Random Node Head Technique and Highest Degree Node Head technique. When number of iterations is set to five, the values are 0.06 and 0.09 for Random and HDNHT respectively and when set to twenty five iterations, the values are 0.14 and 0.17 respectively. These values show that the HDNHT is performing better than the Random technique as higher density value is good for communities.

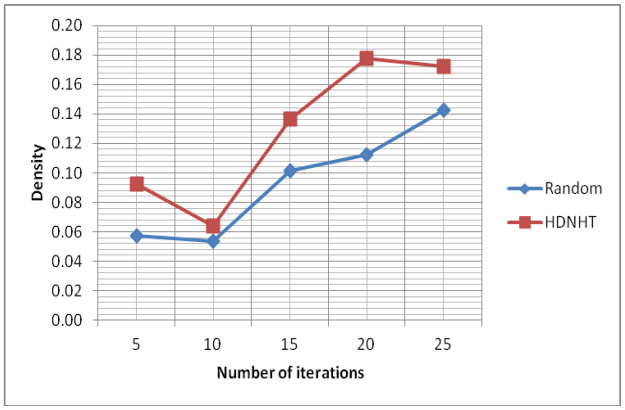

Fig-5: Density Comparison

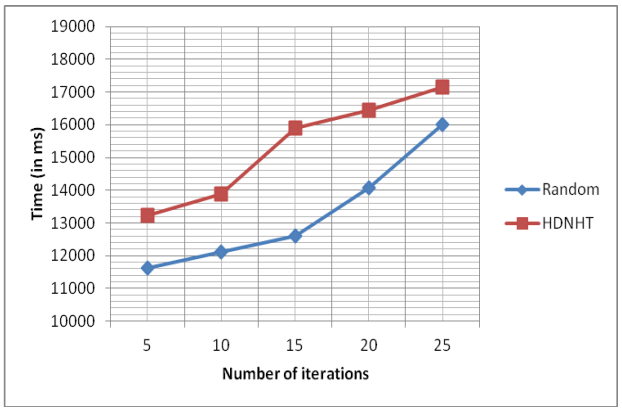

Fig-6: Comparison of Time taken for Execution

The Fig-6 shows the execution time between Random Node Head Technique and HDNHT. In this HDNHT technique is consuming approximately 15% more time than the Random Node Head technique. The values are 11614ms and 13218ms for Random, and HDNHT respectively when number of iterations is five and when set to twenty five iterations, the values are 16014ms and 17143ms respectively.

The Fig-7 shows the memory requirements for each technique. When number of iterations is set to five, the memory requirements are 175.38MB and 175.74MB for Random and HDNHT respectively and when number of iterations is set to twenty five the memory requirements are 190.38MB and 192.54MB respectively. This shows that approximately 3% of the more memory is consumed by the HDNHT technique.

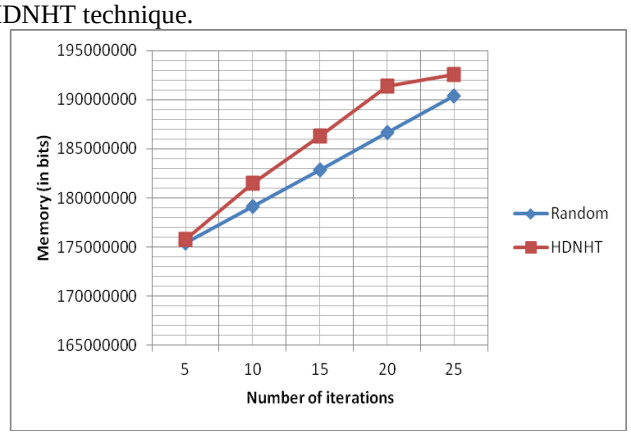

Fig-7: Comparison of Memory Consumption to Execute the Process

V. CONCLUSION

In this paper, the Random Node Head technique and Highest Degree Node Head techniques are proposed to group nodes into communities. In Highest Degree Node Head technique, initially the centroids are chosen randomly and the clustering is done based on the centroids chosen. From the next iteration, the centroids are updated based on highest degree from the clusters formed in the previous iteration. The fitness value is calculated for each set of centroids to choose the best set. The performance of Highest Degree Node Head Technique is compared with the Random Node Head technique in terms of fitness, separability, density, execution time and memory consumption using Wikipedia vote network dataset. The performance comparison shows that in terms of density and separability the Highest Degree Node Head Technique performed well and where as in memory consumption and execution time Random Node Head Technique performed well.

AKNOWLEDGEMENT

Amedapu Srinivas is supported by Technical Education Quality Improvement Program (TEQIP) Phase-II, a World Bank Initiative. The author is thankful to the TEQIP-II and National Institute of Technology, Tiruchirappalli, Tamil Nadu, INDIA for supporting this research work.

REFERENCES

1. Özturk K. Community Detection in Social Networks, Msc. Thesis. Graduate School of Natural and Applied Sciences, Middle East Technical University 2014.
2. McPherson M, Lovin LS, Cook JM. Birds of a feather: Homophily in Social Networks. Annual review of sociology 2001:415 - 444. doi:10.1146/annurev.soc.27.1.415.
3. Watts DJ, Strogatz SH. Collective dynamics of 'small-world'networks. Nature 1998, 393 (6684):440-442. doi:10.1038/30918.
4. Luce RD, Perry AD. A method of matrix analysis of group structure. Psychometrika 1949, 14 (2):95 –116. doi:10.1007/BF02289146.
5. Fortunato S. Community detection in graphs. Physics Reports 2010, 486 (3):75-174. doi:10.1016/j.physrep.2009.11.002.
6. Luce RD, Perry AD. A method of matrix analysis of group structure. Psychometrika 1949, 14 (2):95 –116. doi:10.1007/BF02289146.
7. Fortunato S, Castellano C. Community structure in graphs. In: Computational Complexity: Springer; 2012, 490-512.doi:10.1007/978-1-4614-1800-9_33.
8. Porter MA, Onnela J-P, Mucha PJ. Communities in networks. Notices of the AMS 2009, 56 (9):1082-1097.
9. Danon L, Diaz-Guilera A, Duch J, Arenas A. Comparing community structure identification. Journal of Statistical Mechanics: Theory and Experiment 2005, 2005 (09):P09008. doi:10.1088/1742-5468/2005/09/P09008.
10. Plantié M, Crampes M. Survey on social community detection. In: Social Media Retrieval: Springer; 2013, 65-85.doi:10.1007/978-1-4471-4555-4_4.
11. Kernighan BW, Lin S. An efficient heuristic procedure for partitioning graphs. Bell system technical journal 1970, 49 (2):291-307. doi:10.1002/j.1538-7305.1970.tb01770.x.
12. Newman M. Community detection and graph partitioning. EPL (Europhysics Letters) 2013, 103 (2):28003. doi:10.1209/0295-5075/103/28003.
13. Girvan M, Newman M. Community structure in social and biological networks. Proceedings of the national academy of sciences 2002, 99 (12):7821-7826. doi:10.1073/pnas.122653799.
14. Newman M, Girvan M. Finding and evaluating community structure in networks. Physical review E 2004, 69 (2):026113. doi:10.1103/PhysRevE.69.026113.
15. Newman M. Fast algorithm for detecting community structure in networks. Physical review E 2004, 69 (6):066133. doi:10.1103/PhysRevE.69.066133.
16. Newman M. Analysis of weighted networks. Physical Review E 2004, 70 (5):056131. doi:10.1103/PhysRevE.70.056131.
17. Newman M. Modularity and community structure in networks. Proceedings of the National Academy of Sciences 2006, 103 (23):8577-8582. doi:10.1073/pnas.0601602103.
18. Amedapu Srinivas, R. Leela Velusamy., "Identification of influential nodes from social networks based on Enhanced Degree Centrality Measure", 2015 IEEE International Advance Computing Conference (IACC), pp. 1179-1184, 2015.
19. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008. doi:10.1088/1742-5468/2008/10/P10008.
20. Guimera R, Sales-Pardo M, Amaral LAN. Modularity from fluctuations in random graphs and complex networks. Physical Review E 2004, 70 (2):025101. doi:10.1103/PhysRevE.70.025101.
21. Zhou Z, Wang W, Wang L. Community Detection Based on an Improved Modularity. Pattern Recognition 2012:638-645. doi:10.1007/978-3-642-33506-8_78.
22. Duch J, Arenas A. Community detection in complex networks using extremal optimization. Physical review E 2005, 72 (2):027104. doi:10.1103/PhysRevE.72.027104.
23. Ye Z, Hu S, Yu J. Adaptive clustering algorithm for community detection in complex networks. Physical Review E 2008, 78 (4):046115. doi:10.1103/PhysRevE.78.046115.
24. Wahl S, Sheppard J. Hierarchical Fuzzy Spectral Clustering in Social Networks Using Spectral Characterization. In: The Twenty-Eighth International Flairs Conference; 2015 : 305-310
25. Falkowski T, Barth A, Spiliopoulou M. DENGRAPH: A density-based community detection algorithm. In: IEEE/WIC/ACM International Conference on Web Intelligence (WI); 2007: 112-115.doi:10.1109/WI.2007.74.
26. Dongen SV. Graph Clustering by Flow Simulation, PhD thesis, University of Utrecht. 2000.
27. Pizzuti C. GA-Net: A genetic algorithm for community detection in social networks. In: Parallel Problem Solving from Nature–PPSN X: Springer; 2008, 1081-1090.doi:10.1007/978-3-540- 87700-4_107.
28. Pizzuti C. A multiobjective genetic algorithm to find communities in complex networks. IEEE Transactions on Evolutionary Computation 2012, 16 (3):418-430. doi:10.1109/TEVC.2011.2161090.
29. Hafez AI, Ghali NI, Hassanien AE, Fahmy AA. Genetic algorithms for community detection in social networks. In: 12th International Conference on Intelligent Systems Design and Applications (ISDA): IEEE; 2012 : 460-465.doi:10.1109/ISDA.2012.6416582.
30. Mazur P, Zmarzlowski K, Orlowski AJ. A Genetic Algorithms Approach to Community Detection. Acta Physica Polonica Series A-General Physics 2010, 117(4).
31. Liu X, Li D, Wang S, Tao Z. Effective algorithm for detecting community structure in complex networks based on GA and clustering. In: International Conference on Computational Science (ICCS 07): Springer; 2007:657-664.doi:10.1007/978-3-540-72586-2_95.
32. Tasgin M, Herdagdelen A, Bingol H. Community detection in complex networks using genetic algorithms. arXiv preprint arXiv: 0711.0491 2007.
33. Zadeh PM, Kobti Z. A Multi-Population Cultural Algorithm for Community Detection in Social Networks. Procedia Computer Science 2015, 52:342-349. doi:10.1016/j.procs.2015.05.105.
34. Raghavan UN, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks. Physical Review E 2007, 76 (3):036106.doi:10.1103/PhysRevE.76.036106.
35. Xie J, Szymanski BK. Labelrank: A stabilized label propagation algorithm for community detection in networks. In: IEEE Network Science Workshop (NSW); 2013:138-143
36. Wu Z-H, Lin Y-F, Gregory S, Wan H-Y, Tian S-F. Balanced multi-label propagation for overlapping community detection in social networks. Journal of Computer Science and Technology 2012, 27(3):468-479.
37. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. the Journal of machine Learning research 2003, 3:993-1022.
38. Xin Y, Yang J, Xie Z-Q. A semantic overlapping community detection algorithm based on field sampling. Expert Systems with Applications 2015, 42 (1):366-375. doi:10.1016/j.eswa.2014.07.009.
39. Xin Y, Yang J, Xie Z-Q, Zhang J-P. An overlapping semantic community detection algorithm base on the ARTs multiple sampling models. Expert Systems with Applications 2015, 42 (7):3420- 3432. doi:10.1016/j.eswa.2014.11.029.

40. Xia Z, Bu Z. Community detection based on a semantic network. Knowledge-Based Systems 2012, 26. doi:10.1016/j.knosys.2011.06.014.

AUTHORS PROFILE

Amedapu Srinivas is a full time Ph.D research scholar at NIT Trichy, Tamil Nadu. He did his B.E in Computer Technology from Nagpur University and M.Tech form JNTU Hyderabad. He is been in education field as a teacher for 15 years. His interests includes Social Networks Analysis and Big Data Analytics

Professor Dr. (Mrs.) R Leela Velusamy is an academician of the National Institute of Technology Trichy, Tamil Nadu. She worked as Head of the CSE department. She published several papers in reputed publishers like IEEE, Elsevier, Wiley, Springer, etc. Her interests includes Mobile Ad Hoc Networks, Social Networks, QoS in Routing.