Novel Community Detection and Ranking for Social Network Analysis.

Pujitha Reddy Surapareddy

Abstract—Enterprises are collecting, procuring, storing, curating and processing increasing quantities of Big Data. This is occurring to find new insights that can drive more efficient and effective operations and provide management with the ability to steer the business proactively [1] [2]. Currently, Big Data Analysis blends a traditional statistical data analysis approach with computational ones. Since the dataset types are growing exponentially in popularity, various Community Detection and Community Ranking algorithms are developed for target marketing. Such analyses utilize the concepts of shortest path, closeness centrality, and clustering coefficient.In this study, we developed a community detection algorithm based on centrality and node closeness. We performed Visual Analysis, i.e. the graphical representation of data, to depict an interconnected collection of entities- among people, groups, or products. We also performed Network Analysis (Community Detection and Ranking Algorithms) to analyze the relationships among the entities. The proposed algorithms were applied to multiple datasets and the hidden patterns of each were identified. Among the benchmark datasets, the algorithms were implemented on the American College Football and Karate Club datasets. We were able to predict the next matches, the most popular person in the club, and their relevant connections with a high accuracy as compared to the ground truth. In addition to the high accuracy, it encompasses all the features and predicts the importance of the community leader, which is a key differentiating factor for the propose algorithms. Modularity was used as the metric to compare the effectiveness of the proposed methods with state-of-the-art frameworks. The max modularity of the Jaccard, Cosine, Pearson, Edge Betweenness, and Closeness centrality for the Karate Club with two detected communities were measured as 0.21, 0.23, 0.30, 0.40, and 0.35 respectively. Meanwhile, these values were in the range of 0.58-0.59 for the Football dataset with 12 detected communities.The proposed Community Detection and Community Ranking algorithms outperformed the existing frameworks on scale-free networks. We were also able to identify the hidden patterns of friendships on social media, frequent itemsets purchased together, which can be used to develop recommendation systems to e-commerce portals.

Index Terms—

1 Introduction

Real world networks such as Information, Transportation and Social networks generate voluminous amount of data in today's world [3].Network analysis that involves community detection and ranking has become prominent in predicting the primary features and topology of these kinds of networks. Community is an important parameter that leads to identify connections deep down the network.Ranking can also help in identifying the patterns that influence and cluster similar members of a network in to a community.Ranking determines how important a node is in a Network [4]. There are many ranking measures to find importance of a node in a network such as Degree Centrality, Closeness Centrality, Between-ness Centrality, Eigen Vector Centrality, Katz Centrality, PageRank Centrality. All these centrality measures have their own importance as well as limitations. Each works well only for probing certain phenomena.However in most cases none of them is complete enough to find the overall importance and influence [5]. Degree Centrality is the simplest measure that reflects the number of immediate connections and ignores the influence of those direct links.Between-ness and Closeness centrality are based on the shortest path [6] [7].Although information flow is not always through the shortest path in real world networks.Eigen Vector Centrality overcomes the degree centrality to certain extent by capturing the influence of neighbors.However, it fails if the influence is passed to its neighbors .Katz centrality gives a constant importance to all its neighbors but unnecessarily distribute its importance over the network if there are many outgoing edges to many neighbors [8].In case of undirected graphs, page rank algorithm divides by out degree, which is equal to number of neighbors it have makes the node less important even though if it is important [9]. Eigen vector centrality also makes nodes centrality zero even it have many neighbors and if neighbors degree is zero in sparse graphs [10]. A community (cluster) is a densely connected group of vertices, with only sparser connections to other groups [11]. A network is said to have community structure if the nodes of the network can be meaningfully grouped into sets of nodes such that each set of nodes is densely connected internally [12]. We need a numeric criteria of inside-ness and outside-ness. Something we can try to maximize. A quantity called modularity is often used to decide the best number of communities. [6] The main aim of this research is to create a new clustering and community ranking methods those involve less computations and uses less space and gives efficient clusters. As Community detection is key to understand the structure of complex networks and extract the useful information from Network and this project could be used to find optimal clusters [13].

- *Pujitha Reddy Surapareddy with the Department of Computer Science, California State University, Fresno .*
  E-mail: pujithareddys@mail.fresnostate.edu
-

Manuscript received September 8, 2019

The rest of this paper is organized as follows: "Methodology" introduces the proposed method and its applications. "Analysis" depicts the application of proposed method to real world data sets and comparing them with ground truths. Conclusion and future work discuss the outline of proposed algorithm, its applications and improvements that can be made in future.

2 Literature Review

In this paper [14] a new type of network community in online social networks (OSNs) is identified using the association between network nodes [author] Proposes a virtual community detection algorithms that uses the basics of link prediction to detect communities. Li at [2] says that Link prediction works on a concept called information diffusion. Prime Nodes: If a pair of nodes share two or more consistent neighbor sets, such nodes are called prime nodes.Virtual Communities: The consistent set of neighbors shared by prime nodes along with prime nodes are identified as Virtual communities.The main contributions of this research are: To identify communities, by comparing the similar nodes connected with prime nodes.Link prediction to make the incomplete OSN complete by association algorithms.This research is validated using centrality methods page rank and k-core before and after link predictions in OSN's. The research proposed here in [15] is mainly based on the graph theory.Community is defined as a group with more connections inside than connections outside.There is a possibility that a nodes belongs to more than one community at the same time.This is referred as community overlapping, which we often see in real world networks such as social networks. Community detection starts with the properties of chordal graphs. First,the maximal cliques are identified and it is extend with nodes and edges.Later it is optimized greedily by local clustering function.The proposed algorithm is bench marked using the ground truths of the networks experimented. K-Means clustering is a classical algorithm proposed long back [16]. It always start with choosing appropriate K, which always does not give efficient clustering.It works well for global clusters as it always assigns the nearest items in the radius.If the size of he network is huge, K-Means is always sensitive to initialization.Affinity Propagation is a new type of algorithm that was introduced in this research. Communities are clustered based on nearest neighbor influences a node have. A threshold of nearest neighbors delta is set and influence of a node is calculated.The list is sorted based on the influence and more influenced neighbors are clustered. A semi supervised community detection algorithm is proposed [17] based on the concept of must-links and cannot link.The must-link nodes are started as a seed and it is extended based on the transitive property. Nodes that are similar to must-link nodes make a cluster. Transitive Property does not get hold here in can not links. In the social network, the centrality plays an import role in the graph. Finding the centrality could help to improve the accuracy in classification data mining techniques. Hui[8]studied the correlation of degrees and betweenness centrality to investigate BBS reply networks. And the result showed central nodes with high degree or high betweenness centrality do have high influence and power in online social networks. On the other hand, when central nodes with highest degree and highest betweenness centrality are removed network centralizations decreased typically. [9] provided an idea to use centrality metric such as degree, eigenvector, betweenness and closeness to define the most central state within a country and use this information to design the road/rail transportation networks. [18] presented to use a new metric which namely Cross closeness centrality for measuring the multiplex social network and simple network. The datasets were the families which were from two different areas(Danio Rerio and Florentine). After analyzing, the data showed multiplex networks offers much valuable and concrete information compare to the simple network. Prantik et al. [5] observed the influence users from Twitter and use Degree Centrality and Eigenvector Centrality to collect the data. The result showed that indegree and eigenvector centrality should both be considered when finding users who are influential. [19] presented a new centrality which combined betweenness centrality and Katz centrality to measure the importance of node. This new centrality not only reduced the problem of betweenness centrality which only focused on the shortest path but also solve the problem of Katz centrality which focused on the adjacent nodes. Lingjie[12] gave a new centrality which depended on the betweenness changes caused by the removal of the largest node in the network. This method was useful to identify the functional and structural of importance of the nodes in a network.

3 Methodology

DataSet	Sparse / Dense	Weighted / Un-Weighted
Karate Club	Dense	Unweighted
Dolphins	Sparse	Unweighted
Les miserables	Dense	Unweighted
Football	Sparse	Unweighted
Political Blogs	Dense	Unweighted
Neural Networks	Dense	Unweighted

Clustering coefficient is to know how connected the neighbors are to each other. Mathematically, Clustering coefficient of a node i is

$$CC_i = \frac{2|e_{jk} : V_j, V_k N_i, e_{jk} \in E|}{K_i * (K_i - 1)} \quad (1)$$

where Ni = Neighbor nodes to Vertex j, K=mod Ni is the number of neighbors, Li=number of links between neighbors. If an important node points to many less important nodes unnecessarily other nodes also gets importance, so page rank divides it by out degree which in case of undirected network is not justified. So,we use clustering coefficient and influence as measures and propose a new centrality method. Here we propose community detection based on the common neighbor and influence they have in the network [13]. This influence centrality indicates if two nodes are close to common set of neighbors then they both are closer to each other. This makes the nodes stay in one community. To find the common neighborhood we can make use of the centrality measures. An influence matrix will be generated. First the influence weight for all the nodes

Algorithm 1: Influence and Clustering Centrality

```
Input: Directed Graph , InfluenceLimit
Output: Centrality Matrix
for i ← 1 to n do
    find cc(i)
    Set influence weight of each node to cc[i]
end
for i ← 1 to n do
    Empty open/close list
    Set all nodes to be unexplored
    OpenList-PushStack (Node(i))
    Set Node(i).depth = 0
    while OpenList is not empty do
        currNode = OpenList-PopStack ( )
        Node(i).influence Vector(currNode) +=
         (currNode.depth)2
        Pop all nodes in CloseList with depth
         currNode.depth
        Set those nodes to be unexplored;
        if currNode.depth < influenceLimit then
            for each out-link neighbor j of the currNode
             do
                if Node(j) is unexplored then
                    OpenList-PushStack(Node(j))
                    Set Node(j).depth =
                     currNode.depth + 1
                end
            end
            Set currNode.isExplored = True
            CloseList-PushStack (currNode)
        end
    end
    M(i) = Node(i).influence Vector
end
return M
```

is set to 1. Later based on the path length and the clustering coefficient set the influence for each node. At every level update the influence weight with the clustering coefficient.

4 Results

Karate Club:

In karate club dataset, Node 8 have the CC = 1.0 which means all the neighbor nodes are connected to each other and it forms a clique, where any information easily spreads at this part of network. Node 8 has important nodes in terms of other centralities 1,3 connected to it which means it can easily pass information to other nodes as it forms a clique around it. It is also pointing to only 4 nodes and it justifies the problem that if an important node is pointing many other nodes the other are getting importance unnecessarily. It is also close to many important nodes such as instructor which have high connectedness. The new centrality measure is defined by taking some centralities and ranking methods .so , the central nodes of this ranking includes the nodes in order which have high clustering coefficient, neighbors with high connectedness and this ranking will not give importance which it have unnecessarily by pointing to many nodes.

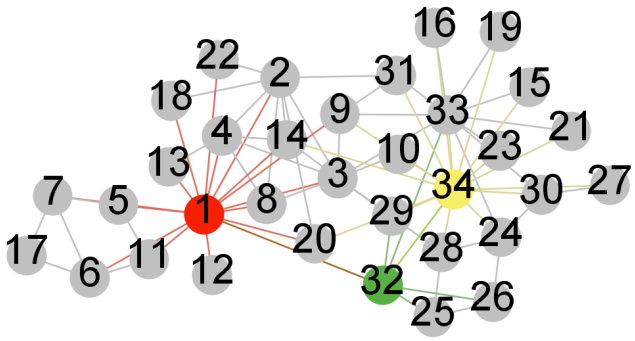

Fig. 1: Community Ranking for Karate Club.

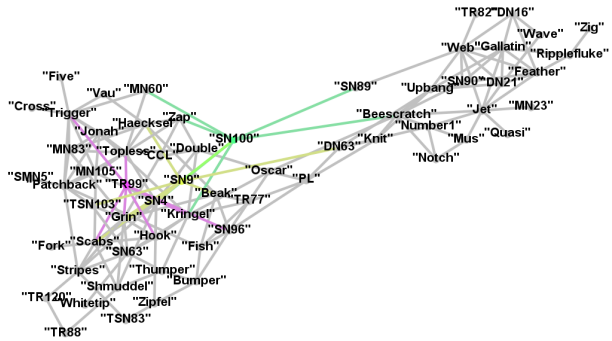

Fig. 2: Community Ranking for Dolphin Community.

In Dolphins Community, there are no nodes with clustering coefficient 1, which means connectedness among neighbors in dolphin community seems less compared to other data sets taken. Notch have the CC = 0.6 which it forms a partially connected graph around it. The neighbors of the Notch are Grin and SN4 which are highly connected nodes. Hook is close to the connected nodes ,have high clustering cofficient , not making less important nodes unnecessarily important.

In football matches conducted, almost all teams that played with walke forest has played with each other.so

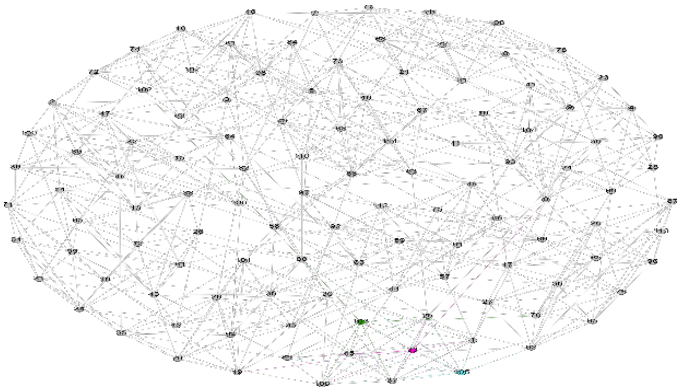

Fig. 3: Community Ranking for Football Data Set .

walke forest team has high clustering cofficient . walke forest played with teams that played highest games in the tournament. It played with very few less popular teams which means a team that played very few times

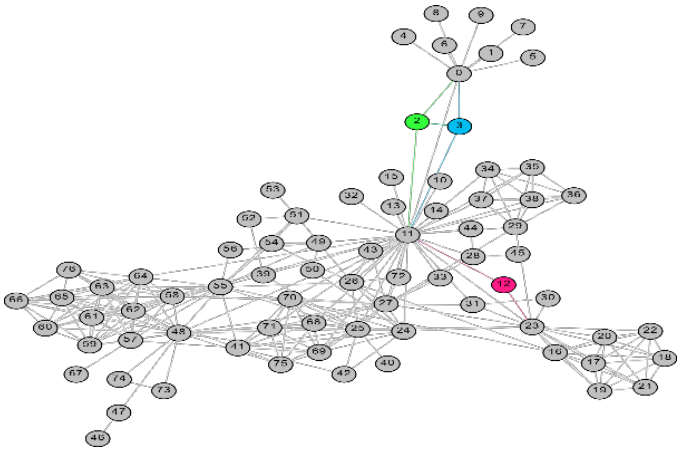

Fig. 4: Community Ranking for Les Miserables dataset.

In les miserables novel Marguerite is the central character according to new centrality measure. That is Marguerite made a good connections i.e., clique around it. It also appeared with most popular character of the novel valijean. It also appeared with Enjolras, Courfeyrac, Fantine which also appeared with important characters.

In les miserables novel Marguerite is the central character according to new centrality measure. That is Marguerite made a good connections i.e., clique around it. It also appeared with most popular character of the novel valijean. It also appeared with Enjolras, Courfeyrac, Fantine which also appeared with important characters.

In figure 5 The clusters are identified with different colors. The nodes with high degree(1,33,34), high closeness(1,3,34) and eigen(1,3,34) are one cluster away (1 is one cluster away from 33,34).The two central nodes 33,34 are in same cluster.so the 4th cluster might be the important community. Homophily seems to be the centrality and closeness.

In dolphins community homophily might be centrality Most of the central nodes discovered by centrality methods

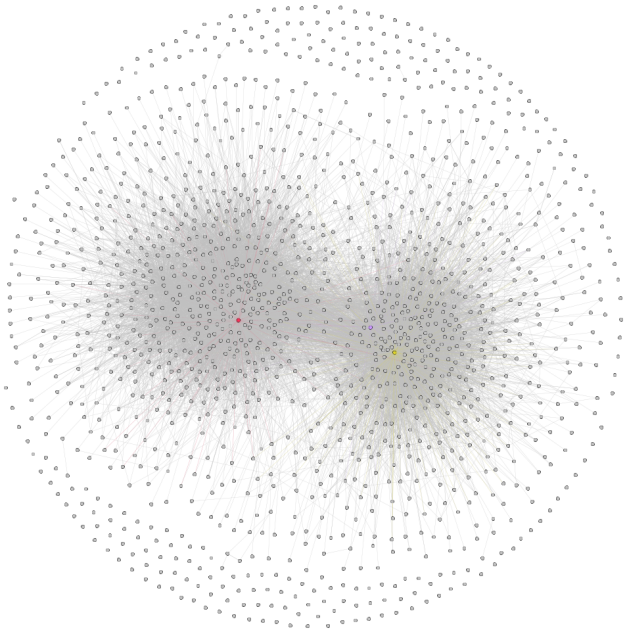

Fig. 5: Community Ranking for Political Blogs.

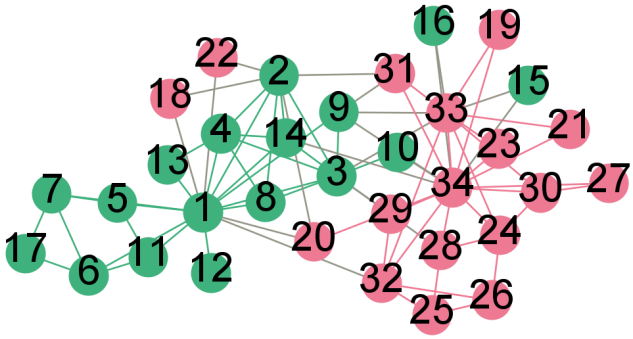

Fig. 6: Visualization of Communities in Karate Club dataset.

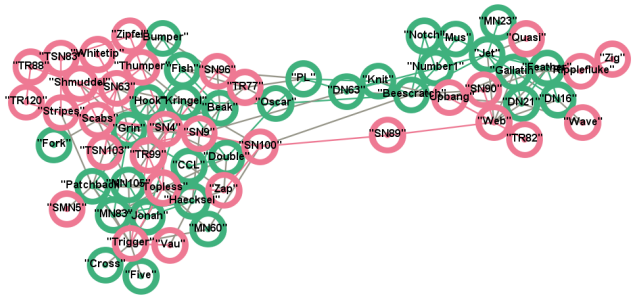

Fig. 7: Visualization of Communities in Dolphins dataset.

are in the clusters 3,4 and 2.This Says those three clusters are the most important communities in the network

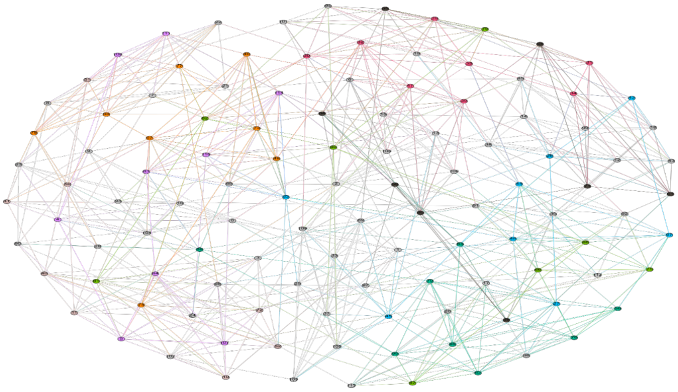

Fig. 8: Visualization of Communities in Football dataset.

In figure 7 The homophily of clusters formed might be geographical location because cluster 1 (Neveda, California, Washington) have all the teams from western states and cluster 2 (Florida, Virginia, Carolina) have all eastern cluster 3(MICHIGAN,ILLINOIS,MINNESOTA) have all the north eastern states cluster 4(TEXAS,OKLAHOMA) have south eastern states There are 10 clusters in the set with maximum modularity.

In figure 8 In this novel all the important characters Gavaroche, Enjolras, Courfeyarac, and Marius belongs to same community which makes the community more central in the entire network. Here also homphily is centrality. There are 4 clusters in the set with maximum modularity

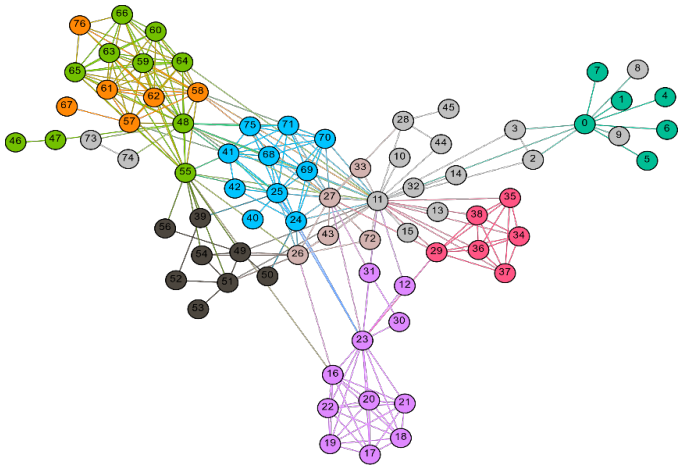

Fig. 9: Visualization of Communities in Les Miserables dataset..

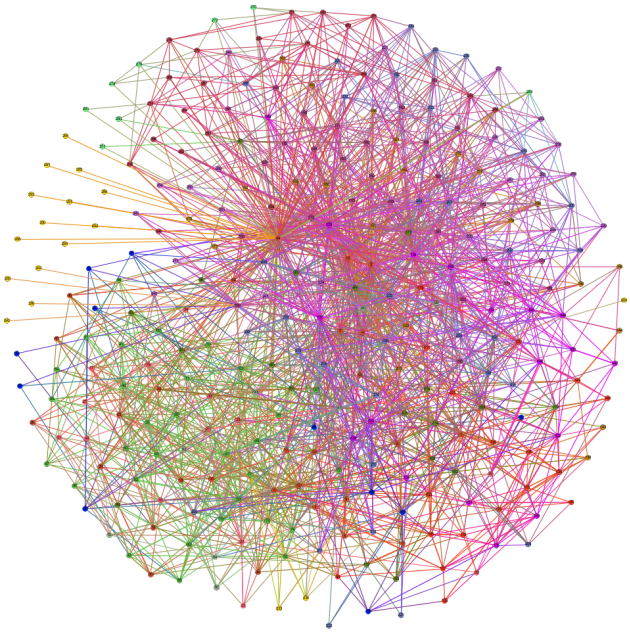

Fig. 10: Visualization of Communities in NueralNetwork dataset..

5 CONCLUSION

Community Ranking: The outcome of this research is deriving a new ranking method by analyzing the existing ones and making new ranking achieve the things that are missing in existing ones [19]. Node 8 is the most important node among all other nodes in the karate club according to new centrality measure in terms of its important neighbors, connectedness among neighbors, it's closeness to all other nodes in the network. In dolphins community which is a social network, Hook is the central node as it is close to highly connected nodes(Grin and SN4) and it formed a good friendship around it which means all friends of hook are friends to each other.

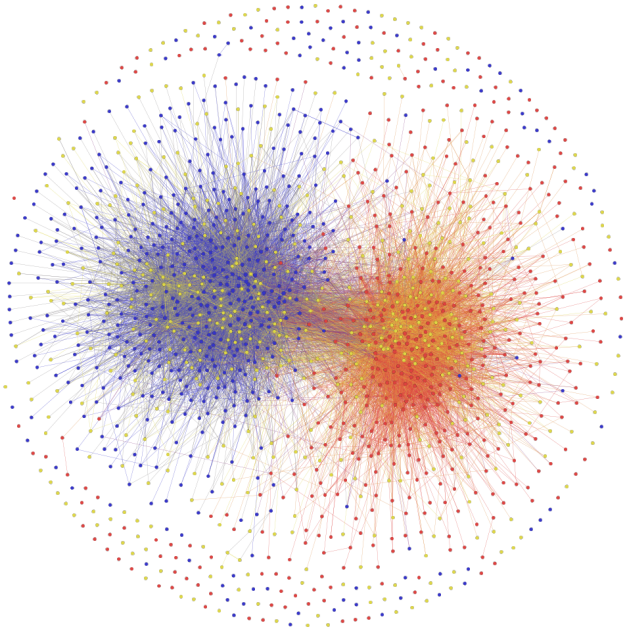

Fig. 11: Visualization of Communities in Political Blogs dataset..

In les miserables novel Marguerite is the important character as it appeared mostly with characters that are highly popular and in football dataset walke forest is the important team as the teams that played with Walke forests are most played teams. Walke forest played with very few teams that are less popular.

Community Detection: The outcome of this experiment is detecting communities by grouping the nodes with high closeness. As a result the dendrogram is formed from the bottom up .After creating the dendrogram modularity is calculated at each level to measure how good the clustering is [20]. The clustering determines the homophily each community have [21]. In karate club the central nodes are distributed over communities where as in lesmis and dolphins many central nodes are together in one cluster.In football dataset purely geographic location is the homophily. Dendrograms and the respective Modularities varied from edge betweenness and Jaccard, Cosine , Pearson Correlation and pathway detection for the karate club, Dolphins ,lesmis data set. But remained same in case of Football dataset. Compared to Edge betweenness the modularity values seems less in Similarity Matrix community detection in all the data sets Except Football.

REFERENCES

[1] Lingjie Zhou, Yong Zeng, Yi He, Zhongyuan Jiang, and JianFeng Ma. Multi-hop based centrality of a path in complex network. In *2017 13th International Conference on Computational Intelligence and Security (CIS)*, pages 292–296. IEEE, 2017.

[2] Zhichao Song, Hong Duan, Yuanzheng Ge, and Xiaogang Qiu. A novel measure of centrality based on betweenness. In *2015 Chinese Automation Congress (CAC)*, pages 174–178. IEEE, 2015.

[3] Matin Pirouz, Justin Zhan, and Shahab Tayeb. An optimized approach for community detection and ranking. *Journal of Big Data*, 3(1):22, 2016.

[4] U Kang, Spiros Papadimitriou, Jimeng Sun, and Hanghang Tong. Centralities in large networks: Algorithms and observations. In *Proceedings of the 2011 SIAM international conference on data mining*, pages 119–130. SIAM, 2011.

[5] Apeksha P Naik and Sachin Bojewar. Tweet analytics and tweet summarization using graph mining. In *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*, volume 1, pages 17–21. IEEE, 2017.

[6] Prantik Howlader and KS Sudeep. Degree centrality, eigenvector centrality and the relation between them in twitter. In *2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, pages 678–682. IEEE, 2016.

[7] Ruchi Mittal and MPS Bhatia. Cross-layer closeness centrality in multiplex social networks. In *2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–5. IEEE, 2018.

[8] Natarajan Meghanathan and Raven Lawrence. Centrality analysis of the united states network graph. 2016.

[9] Alireza Hajibagheri, Ali Hamzeh, and Gita Sukthankar. Modeling information diffusion and community membership using stochastic optimization. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 175–182. ACM, 2013.

[10] Cuijuan Wang, Wenzhong Tang, Bo Sun, Jing Fang, and Yanyang Wang. Review on community detection algorithms in social networks. In *2015 IEEE International Conference on Progress in Informatics and Computing (PIC)*, pages 551–555. IEEE, 2015.

[11] Yanping Zhang, Yuanyuan Bao, Shu Zhao, Jie Chen, and Jie Tang. Identifying node importance by combining betweenness centrality and katz centrality. In *2015 International Conference on Cloud Computing and Big Data (CCBD)*, pages 354–357. IEEE, 2015.

[12] Zhao Yang, René Algesheimer, and Claudio J Tessone. A comparative analysis of community detection algorithms on artificial networks. *Scientific reports*, 6:30750, 2016.

[13] Wenjun Wang and W Nick Street. A novel algorithm for community detection and influence ranking in social networks. In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*, pages 555–560. IEEE, 2014.

[14] Muhammad Sadiq Khan, Ainuddin Wahid Abdul Wahab, Tutut Herawan, Ghulam Mujtaba, Sani Danjuma, and Mohammed Ali Al-Garadi. Virtual community detection through the association between prime nodes in online social networks and its application to ranking algorithms. *IEEE Access*, 4:9614–9624, 2016.

[15] Conrad Lee, Fergal Reid, Aaron McDaid, and Neil Hurley. Detecting highly overlapping community structure by greedy clique expansion. *arXiv preprint*

arXiv:1002.1827, 2010.

[16] Xinquan Chen. A new clustering algorithm based on near neighbor influence. *Expert Systems with applications*, 42(21):7746–7758, 2015.

[17] Jianjun Cheng, Mingwei Leng, Longjie Li, Hanhai Zhou, and Xiaoyun Chen. Active semi-supervised community detection based on must-link and cannot-link constraints. *PloS one*, 9(10):e110088, 2014.

[18] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

[19] Dhanya Sudhakaran and Shini Renjith. Survey of community detection algori ms to identify e best community in real-time networks. *Vol-2, Issue-1*, pages 529–533, 2016.

[20] Andrea Lancichinetti, Santo Fortunato, and Filippo Radicchi. Benchmark graphs for testing community detection algorithms. *Physical review E*, 78(4):046110, 2008.

[21] Vito Latora and Massimo Marchiori. A measure of centrality based on network efficiency. *New Journal of Physics*, 9(6):188, 2007.