International Journal of Pure and Applied Mathematics
Volume 118 No. 20 2018, 745-758
ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue
ijpam.eu

Optimization of Community Detection in Twitter Social Network

G. Suryateja, P.Saravanan

School of Computing, SASTRA Deemed University, India

ABSTRACT

Social networks like Twitter, Facebook are the most visited sites on the internet. These websites have a large amount of data about the people and link between them. Structure of community is the main crucial part of social networks. It has a very wide range of applications in computer science, biology, and social sciences. Community detection will show how the structure of the links will have the impact on the people and the relationship among them. For community discovery, high range of applications hasbeen developed for years and years. Social networks play a major role in the dispersal of innovation and information. Social networks became very famous inthe area of research. In identifying the communities,the proposed system divides the network into regions in the corresponding graph. The persons in the community will be with same qualities and perform similar actions. In this approach, the communities are identified using Twitter streaming data. The twitter authentication keys used for accessing live streaming tweets in the network. Girvan Newman (GN)method which will identify the communities by calculating the betweenness, centrality by estimating the total number of paths which is near in the network. And here we are also comparing our algorithm with some other algorithms to prove that our algorithm is better and efficient after comparing with the other.

Keywords: Community detection, social networks, Betweenness centrality, modularity, similarities among the users.

1.INTRODUCTION

Recently social network analysis have gained more attraction from many communities. It contains nodes and edges. The nodes represent the users, and the edges represent the relationship among the users. The users can be an individual or a group. Many social networks like twitter, face book and LinkedIn have been developed for sharing and finding new contacts and new contents. Social networking sites have gained much growth in the shortperiod. The social network is best represented as a graph. Social network[1] analysis mainly emphasis on analysis of interactions and link between people. It will give a clear as well as mathematical analysis of relationships.it is also same as like analyzing the graph since it is from the topology of the graph the application of social network analysis are network propagation modeling, community-based resource support, aggregation and mining, filtering and social sharing. The relation and actions between peoples in the networks can also identify based on the behavior and position.

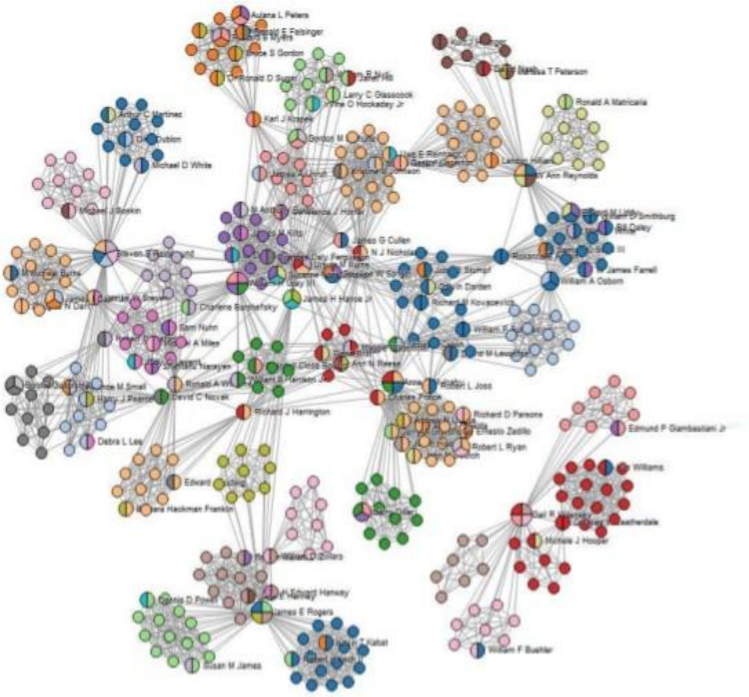

Fig 1.Social network structure.[16]

The social networkalso has an enormous range of interest in various disciplines. The data that is collected from Twitter tweets can also be modeled like as a graph. The Fig [1] will give us a clear idea that how the actual structure of a social network will appear. Though thereare nothing strict rules that we should extract the structure of the community. The community may be synonymous with clusters, groups in different contexts. The structure of the community is one of the great future in the social networks. The graphs that will generate randomly doesn't have such type of feature. And the main aim is to divide the networks and to form the communities, where each node is assigning with separate community[2]identity. It is related to the clustering algorithm.

This method is based on the live streaming data that is obtaining from Twitter.Thesimplicity that is present in our algorithm will make to extract the structure of the community from a large number of social networks along efficiency. They also tend to divide networks into many modules in any cases. And our algorithm will never get a division if there is no structure of the community is identifying. It will alt when there is no longer smaller scale of communities structure is found. The graph analysis which we will use while finding the communities has become very important in understanding the complex networks. Mostly the many of the complex systems will be represented as complex networks where the vertices will denote the elements present in the system,and the links will denote the communication between them.

The social networks will play the major role in the identification of the diffusion of the corresponding information, the ideas and also the innovations using this advantages the subject of the several parts that have moved through the networks to achieve the goals. The main and the major part is to detect the community structure.

The discovery of the communities in social networks is the first and foremost problem in the network of science. The community detection goal is to find the clusters as the subgraphs in the given networks. The community is a group of that belong to the same group. The way in which the nodes are organized is playing a vital role in the process of spreading. The study about the role models will help us to understand better about why some trends will be adopted very fast than others. Also,it helps the advertisers and the marketers in designing more effective campaigns.

2. RELATED WORK

In identifying communities, several methods are proposed. This method will work with a variety of algorithm like data mining, graph mining algorithm to detect task of communities.Fig.2 shows a brief idea about the formation of community structure.Some other proposed methods are sub-detection of community and detection of communities in the aspect of the web-scale network, detection of social networkarchitecture and internet of things, detection of weighted networks.

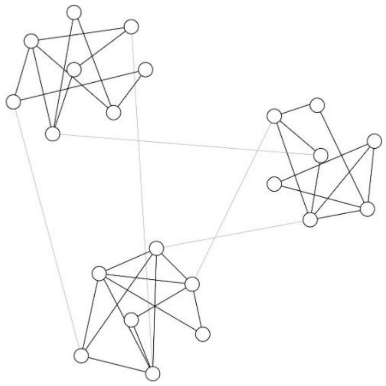

Fig 2. Community structure[17]

New-Girvan algorithm is introduced for detecting of sub-community and community detection. It has two varity features: first, uses iterative removal of edges for splitting to communities, edges removal can be identified by using “betweenness,” second, the measures will cruciallycalculate again later every removal of edges.The sub-community contains two or more nodes for distributed community detection the author introduced a format for finding community partitions using billions of edges.

The methods of identifying the communities is an integration ofInternet of Things (IoT)[3] and also social network architecture which works with graph mining. This technique considers mutual friends as advantages for advising friends, and suggestions should be basedon some mutual friends. Community detection [4] for weighted networks will use clustering method; the main motive is to maximize overall weight for all chosen clusters and maximize the same features among the chosen clusters. Here overall weights of all chosen clusters are calculated with the similarity between the clusters.

The community detection research is forcharacterizing strong social group by network property which supports in understanding the structure of social networks[5] as well as the function of the social network. It will alsohelpin future network growth. The methods to discover the communities are all based on the clique percolation methods, the agglomerative clustering. All these techniques will emphasizeonly on the graph structure and their communities,but it does not consider the user interest and also the interactions. And also the effect the user's fame on the networks.

Some techniques will not allow registering the users in other communities; a few researchers had also believed that in some of the applications, every node wouldbelong to one community, but most of the applications need nodes for overlapping[6] with each other. As a challenge of solving this method the researchers had introduced the Bayesian probability model and this probability model the community members to overlap with others, and this technique mainly focuses on the graph network.

The social networkcanbe partitioned as static as well as dynamic properties. The static properties are viewed with snapshots of the particulargraphand dynamic properties where viewed by the structure of the network. These properties are either weighted or unweighted graphs, weights denotes multi-edges (e.g., several tweet messages from Twitter for a single person at one time, Different authors gave different definitions for community detection[7]. The research that is done in community detection is in several ways and based on different ideas. There are lots of many other challenges are there which are in process to cover, like scalability issue.

The main advantages of the detection are accessing the information.All these tasks are same as the data mining problem which is present in the network analysis where all the traditional mining algorithms will return only the single variant pattern structure along with the support. In the discovery of the community, we will find only one important structure and the expected list of a group of the vertex which constitutesthe structureof the network. Also, a past technique for this definition can also be identifying in the family of the block model solutions. Where particularly some of the work will focus only on the structural equivalence and its definition. With depending upon the structure and the analysts will also find the communities which do not overlap with any one of the before categories. This strategy will work incrementally.

This approach is mainly motivated using the Girvan New man algorithm [9] to estimate the similar sharing service networks, the factors of favor volume and the flicker social networks. The favor volume means thepicture that is liked by the other members. People who are having most influenced fans will have the greater weights compared to any other factors. A person with more favor converge, the probability that is occurring more recently will have high validity compared to others along with the skill in the networks

3. PROPOSED APPROACH

We proposed an advanced approach for identifying the communities in the social networks. In this method we will mainly focus on the Twitter data where we are collecting the information from the live streaming tweets by using twitter API[10]. The user have been considered as nodes and his friends have been considered as edges using the keywords which arerequired (e.g., cricket team) and should consider the friends who are following him frequently For forming communities. For identifying the communities the following steps to be considered and these depicted in Fig.3.

1. Collecting the data from the twitter
2. Pre-processing the collected data if required
3. Creating the individual users
4. Identifying the friends of the users
5. Overlapping the friends with the users
6. Forming community structure

- Step(1) here we have to collect the required data which is usingthe twitter based on the live streaming tweets with the help of the Twitter API.
- Step(2) here we have to preprocess[11] the collected dataset to provide input about the users and also his friend's table according to the algorithm if necessary

- Step(3) here we have to select a count of users (e.g., 10) and have to create separate individual clusters for each user named as nodes.
- Step(4) here we should identify the friends for eachuser who are related to them and also interrelated (edges)
- Step(5) in this step we have to connect or overlap the individual group of friends list to individual users how are related to them and form a small community(nodes to edges)
- Step(6) after applying the required algorithm we have to form them as a large community structure

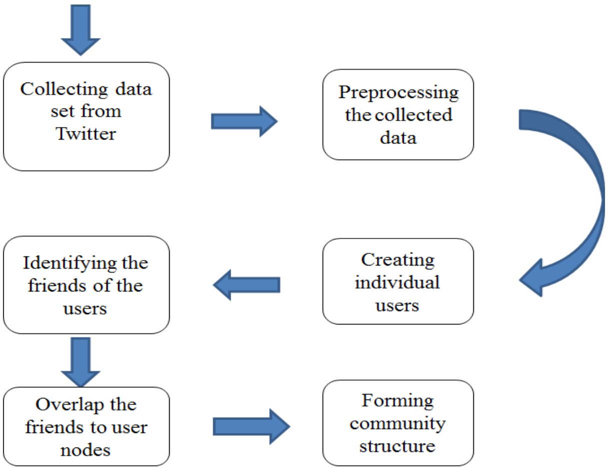
Collecting data set from Twitter
Preprocessing the collected data
Identifying the friends of the users
Creating individual users
Overlap the friends to user nodes
Forming community structure

Fig 3.Community detection identification steps in proposed method

4. Methodology

4.1.Girvan Newman algorithm

This is one of the finest methodswhichare usingto identify the communities in the social networks. It will determine the communities by continuously removing the edges from the original networks which are not relevant to that particular user and forms communities.This algorithm mainly focuses on the concept called *"edge betweenness"* which is used for determining the communities in huge and also complex networks. Here we have to consider the edges which are very nearer and should make it asthe most bounded or "*between*' communities. The count of shortest paths between the pair of enclosed nodeshas been calculated. The edges which are having high *"betweenness"* scorehave been avoided. The procedure has been repeated till a single node is derived.

4.2. Algorithm procedure:

- First, we have to calculate the *betweenness* score of all the existing edges in the network. It can be calculated by estimating the count of shortest paths enclosed bya pair of nodes.
- Next step is to remove the edges which are having the highest *betweenness* score among the nodes in the network
- The *betweenness* of all the edges affected after removal is recalculated.
- Should repeat the process until the edges with more *betweenness*are removing.

- This removal will help in decreasing the running time of the process.

Betweenness centrality

$$BC(V) = \sum_{u,v\in v} \left(\frac{\sigma_{uw(v)}}{\sigma_{uw}}\right)$$

σ_{uw} = Total count of shortest paths among node *u* and *w*

$\sigma_{uw}(v)$ = Total count of shortest paths among node *u* and *w* which passes through *v*

Here $\sum$ is the summation of all the nodes and its edges

V- particular vertex node for which we are calculating betweenness and centrality (if the flow pass through that number of shortest paths we need to count)

5. Test case

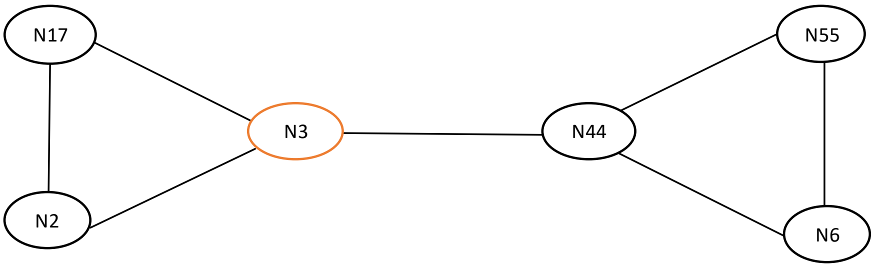

Fig 4: A case study for identifying the betweenness centrality.

Using text case for identifying the betweenness centrality for node *"C"*while performing each action we if it passes through *"C* " note the number of shortest paths it contains else leave it as 0.

Table 1. Betweenness Calculation for selected paths

Selected path	σ_{uw}	$\sigma_{uw}(v)$	$\sigma_{uw}(v)/\sigma_{u}$
(N2,N4)	1	1	1
(N2,N5)	1	1	1
(N2,N6)	1	1	1
(N4,N5)	1	0	0
(N4,N6)	1	0	0
(N1,N2)	1	0	0
(N1,N4)	1	1	1
(N1,N5)	1	1	1
(N1,N6)	1	1	1
(N5,N6)	1	0	0

Table 1. shows the various betwenness values calculated for selected paths in connected networks shown in Fig 4. Here the node *C* and *D*

are atthe same level,so the total betweenness centrality of node ***"C"*** is =6 as mentioned in the above diagram.

6. Connected Components

Strongly connected components (SCC) are the maximum set of vertices such that every vertex is reachable from every other vertex.

- Used in various applications of the networks
- Platform for constructing social relations that forms communities with similar interests and activities.
- The strongly connected components can be applied for the social networks graph to identify smaller groups of nodes, representing users, related to each other by some specific criteria and advertise the onlyto the groups which form the target audience.

$$G(V/E)$$

Here V= vertices, E = edges, G= strongly connected components in the subgraph.

7. Girvan Newman optimization method

In this, we have to divide the network and should calculate the modularity measure.

1. Divide the given network:
   After: dividing the network as shown in Fig 5 we have to calculate the modularity (Q) using the formula given below.

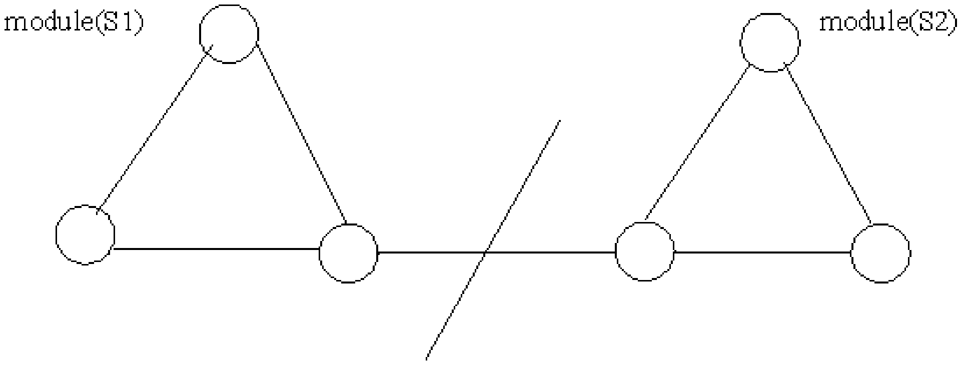

Fig 5: Identifying the connected components

After dividing the network, we have to calculate the modularity (Q) using the formula given below.

$$Q=\sum\left(\begin{pmatrix}\text{Observed fraction}\\ \text{of link in graph}\end{pmatrix}-\begin{pmatrix}\text{Expected fraction}\\ \text{of links in the group}\end{pmatrix}\right)$$

Here we have to find the number of links that are passing through each nodeand also the degrees in the network using the formula.

2. Calculate the modularity(Q)using the below-given formula.

$$Q=\sum_{s=1}^{nm}\left[\frac{\gamma s}{L}-\left(\frac{ds}{2L}\right)\right]$$

Here L= The total count of links of the network.

n_m= Count of modules in the network.

d_s= additionof the degrees of nodes in modules *"s"*.

Finally,here we should have to combine or add both the values of the modularity.For this calculation processes Fig[5]used for identifying the components which are connected to the network. we should identify the total number of links between each node and also degrees in that corresponding network.

8. Dataset

Here we are identifying the communities by using Twitter as a dataset with the help of the live streaming tweets. In this proposed approach we are using the Pakistan super league (PSL) as a community using the # tags that were present in this particular community,and we are going to identify the hidden communities the processes will be as follows

1. Here we have to identify each and individual # tags that were present in the community.
2. Next, we have to identify the users who are performing tweets on this particular # tags.
3. Then we have to find which user is performing tweets on particular # tags.
4. Finally, we have to form individual community structure for each # tag.

9. Comparative Analysis

As a comparative analysis here we are comparing the Girvan Newman algorithm with some other algorithms like *structural algorithmby theRosvall and Bergstrom,Spectral algorithm[12]*by *Donetti and Mu~noz* here in the *Rosvall and Bergstrom* algorithm the identification of the optimal cluster for the purpose better compressing the data on the construction of graph structure.

9.1. Spectral algorithm

Input: the related matrix $S\epsilon\mathbb{R}^{n\times n}$ count k for clusters to be build up.

- Create a same related graph with any of the ways as explained let W be as the adjacent weighted matrix.
- Perform calculation for the unnormalized Laplacian L.
- Calculate the initial N eigenvectors a_1,………,A_k of L.
- Let $U\epsilon\mathbb{R}^{n\times k}$ as matra ix having the vectors having inter related a_1,…….,a_k as columns.
- For i=1,……..,n, here $y_i\epsilon\mathbb{R}^k$ is anglthe e that is inter related to the row I of U
- Batch the points $(i_i)_{i=1,\dots,n}$ in$\mathbb{R}^k$using k-means algorithm to form clusters C1,…..,Ck.
- Output: clusters s1……sk with $A_i = \{j|y_j\epsilon C_i\}$

So it is only possible to see the real structure of the network only after cracking the compressing the date,and in case of *Donetti and Mu~noz[13],* this also belongs to one of the hierarchical clustering methods where it will consider the eigenvector components. The result for the usage of eigenvectors in the network will be inferred by identifying the link between the eigenvectors and detraction of quadratic form

$$\sum_{links}(x_i - x_j)^2 = x^T Lx$$

9.2. DBSCAN Clustering algorithm

The DBSCAN algorithm [14]is one of the clustering algorithmswhere we can identify the clusters within the other clusters here initially we should arrange the data which we are going to perform cluster operation, and the length among the tow should be consider as our next input parameter and the next input will be used for identifying the sense about the clustering algorithm, and this will clearly show us whether the points where same or different and finally we should mention another sensitivity component that will identify if we should start another different cluster to the given data point.

Algorithm

Let K={k1,k2,k3,………kn}are the batch of data points. And the DBSCAN main need two paramedics: Ɛ(eps) and also smallest count of points where needed to model a cluster (minutes).

1) Start from a new point which is not visite yet.
2) Next,collector identify the neighborsto the point by using Ɛ (which means the points which were present around the Ɛ are the neighbors for that point).
3) If the neighbors over this point aresufficient,then the clustering action will begin and should mark the point if it is conversed or visited if not we should mark the point as noise(in future this particular point can again turn anan element in the cluster).
4) When a point is identifify as an element in the cluster, then its corresponding surrounding points is also an element in the cluster is identified.
5) Next one new point which is not visit should be retrieved and refined.
6) This procedure will keep on going up to all the points where considered as visited.

As a result, the DBSCANalgorithm is not good in handling the changing density clusters,and it also fails in calculating the neck based datasets,and it will not work efficiently for Data with high dimensional functionalities.

9.3. Girvan –Newman algorithm

- The Girvan Newman algorithm (GN)[14] is one of the better algorithms for identifying the communities.
- Basic assumption:
  - ➢ Calculate the betweenness centrality of every edge.
  - ➢ Now separate the edges having maximal scores.
  - ➢ Re-calculate all remaining scores
  - ➢ Continue the step 2 process.
- time complexity: O (n^2)
- a lot of deviations have proposed to reform attention by applying by using various betweenness measures.

This algorithm is one of the finest and fastest models with the worst case time O(n^3) and O(n^3) for the sparse graph. As a result, the Girvan New man algorithm is very better and also efficient method when compared to another algorithm that were discussed above.

10.Result

After performing several types of techniques on community detection on Twitter data using the Pakistan super league (PSL), the communities are identified,and the status of the result is as follows.It mainly consists of two steps. Before the clustering processes for each and individual # tags the community for the corresponding (PSL) will be as follows in Fig 6.

10.1. Before clustering

Here we are forming the community with the help of the data which have been extract from the twitter # hash tags,and it will show all the users who are performing tweets on the # tags that spresent in the community.

10.2. After clustering

Here we have to form the individual clusters from the original community by considering the users who areperform the tweets on the particular # tags,and it will be as follows as shown in the below figure.

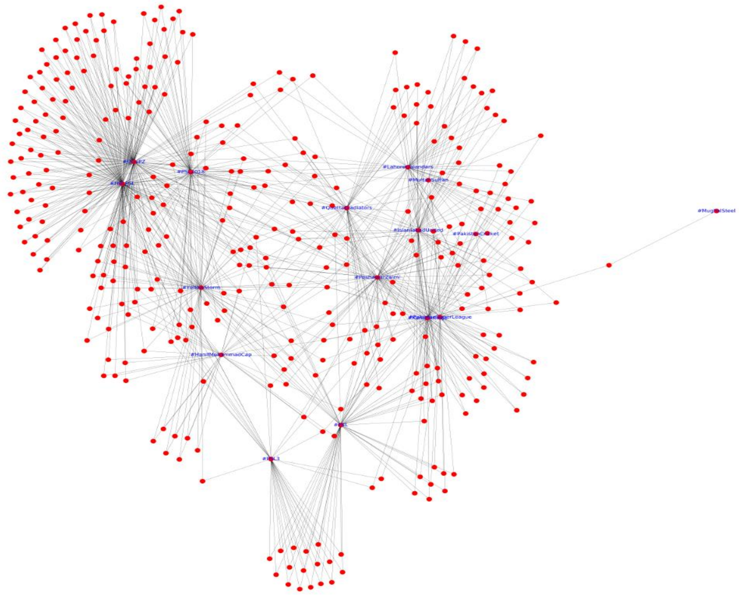

Fig 6.Before clustering graph

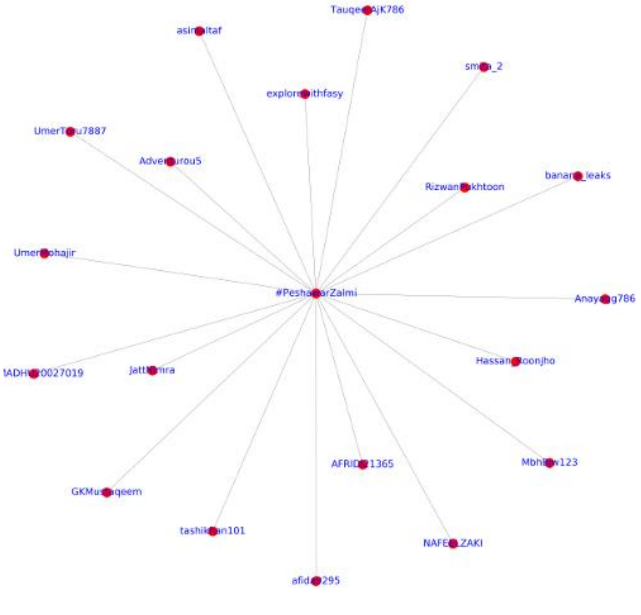

Fig 7.Community structureafter clustering for individual # tag (1).

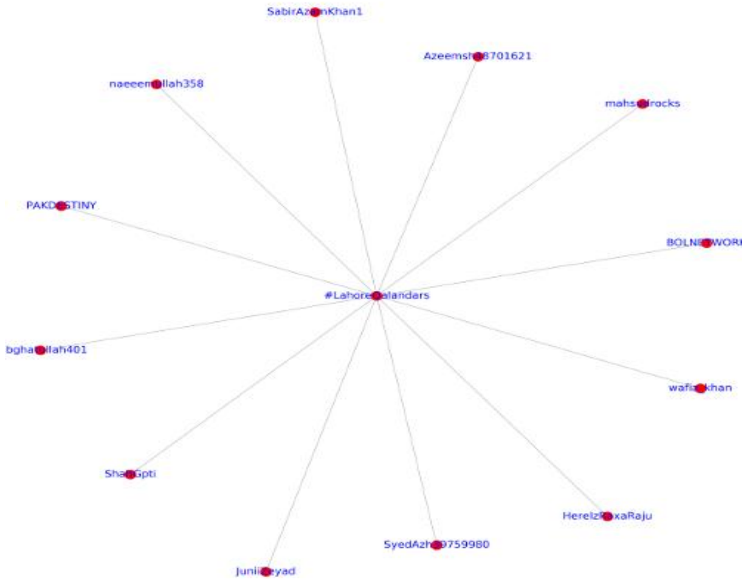

Fig 8.Community structure after clusteringindividual # tags(2)

Fig [6] shows us about the overall community structure before clustering in to individual communities.

Fig [7] shows us about the community structure after clustering for each individual hash tag.

Fig [8] shows us about the community structure for second individual hash tag.

11. Conclusion

Communities will play a major role in the visualization of the huge social network that can also help for understanding the actual structure of social networks. The paper describes various types of algorithms and concepts which is related to detection of the social network community.In proposed approach along that applications among types of a network of the community detection (i.e. twitter dataset) is also properly discussed. And a dynamic network which is consider as a special type of network which contains the connected transactions which have the repeated interactions. And also there is a large number of studies in exploring the online networks. Whenever the updating of the information is completed and also the information about actions which is performed by the users, with minimum of the time complexity and with the local processing by the help of the maintenance of the Girvan Newman (GN) algorithmalso compared our algorithm with many other algorithms and proved that our algorithm is best, efficient compared to other algorithms that where mentioned above.In this article, the proposed method identified the community structure in the social networks using the twitter data,and it will also be help for the researchers who will work in identifying the communities.

References

[1] Choudhury, D., Bhattacharjee, S., & Das, A. (2013, September). An empirical study of community and sub-community detection in social networks applying sthe Newman-Girvan

algorithm. In *Emerging Trends and Applications in Computer Science (ICETACS), 2013 1st International Conference on* (pp. 74-77). IEEE.

[2] Xiong, Z., & Wang, W. (2009). Community detection in social networks employing component independency. *Modern Physics Letters B*, *23*(17), 2089-2016

[3] Misra, S., Barthwal, R., & Obaidat, M. S. (2012, December). Community detection in an integrated Internet of Things and social network architecture. In *Global Communications Conference (GLOBECOM), 2012 IEEE* (pp. 1647-1652). IEEE.

[4] Shen, K., Song, L., Yang, X., & Zhang, W. (2010, October). A hierarchical diffusion algorithm for community detection in social networks. In *Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2010 International Conference on* (pp. 276-283). IEEE.

[5] Ahajjam, S., El Haddad, M., & Badir, H. (2015, November). LeadersRank: Towards a new approach for community detection in social networks. In *Computer Systems and Applications (AICCSA), 2015 IEEE/ACS 12th International Conference of* (pp. 1-8). IEEE

[6] Wang, Z., Zhang, D., Zhou, X., Yang, D., Yu, Z., & Yu, Z. (2014). Discovering and profiling overlapping communities in location-based social networks. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *44*(4), 499-509.

[7] Donetti, L., & Munoz, M. A. (2004). Detecting network communities: a new systematic and efficient algorithm. *Journal of Statistical Mechanics: Theory and Experiment*, *2004*(10), P10012.

[8] Wang, Z., Zhang, D., Zhou, X., Yang, D., Yu, Z., & Yu, Z. (2014). Discovering and profiling overlapping communities in location-based social networks. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *44*(4), 499-509.

[9] Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences*, *99*(12), 7821-7826.

[10] Moosavi, S. A., & Jalali, M. (2014, February). Community detection in online social networks using actions of users. In *Intelligent Systems (ICIS), 2014 Iranian Conference on* (pp. 1-7). IEEE.

[11] Moosavi, S. A., Jalali, M., Misaghian, N., Shamshirband, S., & Anisi, M. H. (2017). Community detection in social networks using user frequent pattern mining. *Knowledge and Information Systems*, *51*(1), 159-186

[12] Donetti, L., & Muñoz, M. A. (2005, July). Improved spectral algorithm for the detection of network communities. In *AIP Conference Proceedings* (Vol. 779, No. 1, pp. 104-107). AIP.

[13] Hsu, D., Kakade, S. M., & Zhang, T. (2012). A spectral algorithm for learning hidden Markov models. *Journal of Computer and System Sciences*, *78*(5), 1460-1480

[14] Khatoon, M., & Banu, W. A. (2015). A survey on community detection methods in social networks. *Int. J. Educ. Manage. Eng.(IJEME)*, *5*(1), 8.

[15] Bickel, P. J., & Chen, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proceedings of the National Academy of Sciences*, *106*(50), 21068-21073.

[16] https://bulaza.wordpress.com/2011/11/29/orgpedia-board-members-

[17] network/Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences*, *99*(12), 7821-7826.