IJCTA, 9(22), 2016, pp. 105-113
© International Science Press

Performance Analysis of Social Network Algorithms on Real World Networks

Bhavna Arora* Archana Singh Gaurav Dubey*** Neha Gaur******

Abstract : In the era of social networking, the size of networks is increasing day by day. Since the nodes in a network may contain some hidden information, it is necessary to study the community structure. In the literature, there are several existing community detection algorithms to detect the community structure algorithms like Walktrap, Multilevel, Eigenvector and InfoMap algorithm etc. In this paper, we have analyzed the modularity value of each of the algorithms of eight dissimilar sized real networks. We have also investigated the community count of networks using each of the algorithms. The performance of different algorithms has been analyzed by using different performance measures and we have emerged with an analysis that the multilevel algorithm works best among all the other algorithms and the Variance of Information (*vi*) is found to have dissipated the best value among four other performance measures. All the analysis is performed by using R programming language. The plot of some real-world networks has also taken in the R- studio. The real networks are analyzed in terms of modularity as well as time complexity.

Keywords : Real world networks; algorithms; community detection; performance measure.

1. INTRODUCTION

The networks are designed to be the natural way of representing the interaction of several nodes via links. In a social graph, the nodes that bear the denser intra connections are said to possess the community structure. This community structure has been studied in the literature which leads to the several community detection algorithms. Some of the algorithms are based on compression techniques, modularity based techniques, similarity based techniques etc. with an aim of detecting the community structure. The Walktrap algorithm[5], being a similarity based technique, uses the hierarchical approach to unearth the community structure, whereas the Eigenvector algorithm is a spectral-partitioning approach that uses the modularity matrix. Two other algorithms, namely, Multilevel and InfoMap are used in large-sized graph to detect the structure of the community. There exist several ways of detecting such communities, but we have used a different approach by using R-programming language. The R - environment is defined to have a complete package for calculation of data and possess high storage and data handling facility. It was ensued at Bell's laboratory by Rick Becker, John Chambers and Allan Wilks. It provides various facilities of writing R- scripts, functions, commands and other statistical features. The R- programming has found to be advantageous in terms of efficiency and ease of use. Many cran packages are supported by R-programming language like igraph, bigdata analytics, R DataMining package. Here, we have used igraph package-a network analysis package, open source software that is programmed in Python, C/C++ and R-programming language. The extant research explored the similar types of social communities' structure using various approaches and algorithms. In the past, all the algorithms tested and used were applied on limited metrics. Since, all the

* Department of Information Technology Amity University, Uttar Pradesh, Noida, India arora.bhavna244@gmail.com
** Department of Information Technology Amity University, Uttar Pradesh, Noida, India archana.elina@gmail.com
*** Department of Information Technology Amity University, Uttar Pradesh, Noida, India gdubey1977@gmail.com
**** Department of Information Technology Amity University, Uttar Pradesh, Noida, India nehagaur199391@gmail.com

approaches belongs to different categories, it was difficult to attain summarized results. The state-of-art shows that that no investigation has been made yet, on the real time social networks. The different algorithms were used in the past but they were lacking in the analysis on real world data sets. In this paper, the parameters and the algorithms used for the comparisons focused more on real world networks to detect overlapping communities. In this paper, it explored the response of real time social network using the significant metrics. The increasing size of networks compels the researchers to analyze the information presented in them. Thus, the study of community structure has become important to understand. Since, the community is based on the concept of dense connections between the nodes [13], the literature explored several existing community detection algorithms like Walktrap algorithm, Eigenvector algorithm, Multilevel algorithm, FastGreedy algorithm, EdgeBetweenness algorithm, InfoMap algorithm etc. We opted for some commonly known community detection algorithms that are known for their robust performance. Some algorithms like EdgeBetweenness [14] are so prolonged that they are ignored in the experimental analysis.The organization of the paper is as follows- In this paper, we studied the quality of partition, *i.e* the modularity value obtained by using each of the algorithms. The section 1 describes the overview of community detection algorithms, namely, Walktrap algorithm, Eigenvector algorithm, Multilevel algorithm and InfoMap algorithm. The section 2 illustrates the performance measures, *i.e* nmi, *vi*, rand index, adjusted rand index and split-join and modularity followed with experimental analysis in section 3. The section 4 consists of results and discussion. The article concludes with the conclusion in section 5.

2. PROPOSED ALGORITHM

The networks are designed to be the natural way of representing the interaction of several nodes via links. In a social graph, the nodes that bear the denser intra connections are said to possess the community structure. This community structure has been studied in the literature which leads to the several community detection algorithms. Some of the algorithms are based on compression techniques, modularity based techniques, similarity based techniques etc. with an aim of detecting the community structure. The Walktrap algorithm, being a similarity based technique, uses the hierarchical approach to unearth the community structure , whereas the Eigenvector algorithm is a spectral-partitioning approach that uses the modularity matrix. Two other algorithms, namely, Multilevel and InfoMap are used in large-sized graph to detect the structure of the community. There exist several ways of detecting such communities, but we have used a different approach by using R-programming language. The R - environment is defined to have a complete package for calculation of data and posses high storage and data handling facility. It was ensued at Bell's laboratory by Rick Becker, John Chambers and Allan Wilks. It provides various facilities of writing R- scripts, functions, commands and other statistical features. The R- programming has found to be advantageous in terms of efficiency and ease of use. Many cran packages are supported by R-programming language like igraph, bigdata analytics, R DataMining package. Here, we have used igraph package- a network analysis package, an open source software that is programmed in Python, C/C++ and R-programming language.The extant research explored the similar types of social communities' structure using various approaches and algorithms. In the past, all the algorithms tested and used were applied on limited metrics. Since, all the approaches belongs to different categories, it was difficult to attain summarized results. The state-of-art shows that that no investigation has been made yet, on the real time social networks. The different algorithms were used in the past but they were lacking in the analysis on real world data sets. In this paper, the parameters and the algorithms used for the comparisons focused more on real world networks to detect overlapping communities. In this paper, it explored the response of real time social network using the significant metrics.

The increasing size of networks compels the researchers to analyze the information presented in them. Thus, the study of community structure has become important to understand. Since, the community is based on the concept of dense connections between the nodes [13], the literature explored several existing community detection algorithms like Walktrap algorithm, Eigenvector algorithm, Multilevel algorithm, FastGreedy algorithm, EdgeBetweenness algorithm, InfoMap algorithm etc. We opted for some commonly known community detection algorithms that are known for their robust performance. Some algorithms like EdgeBetweenness [14] are so prolonged that they are ignored in the experimental analysis.The organization of the paper is as follows- In this paper, we studied the quality of partition, i.e the modularity value obtained by using each of the algorithms. The section 1 describes the overview of community detection algorithms, namely, Walktrap algorithm, Eigenvector algorithm, Multilevel algorithm and InfoMap algorithm. The section 2 illustrates the performance measures, i.e nmi, vi, rand index, adjusted rand index and split-join and modularity followed with experimental analysis in section 3. The section 4 consists of results and discussion. The article concludes with the conclusion in section 5.

2.1. Community Detection Algorithms

Walktrap algorithm : This algorithm devised a distance measured being introduced by Pascal Pons and Latapy. It used a hierarchical clustering approach efficiently to procure the community structure. Consider a graph with '*n*' communities and randomly chose a vertex, say '*x*'. Now, from vertex *x*, calculate the distance between all the adjacent neighbors. Then, the partition P = {{*x*}, *x*∈V} of a graph into communities follows the three steps

1. Choose the two communities to be united based on the distance measure.
2. Integrate the two communities to form a new one and generate a new partition.
3. Update the distance between them

2.2. Eigenvector algorithm

This algorithm was a kind of spectral-partitioning algorithm which used a modularity matrix for detecting communities. It was well explained by M.E.J Newman[8], who expressed the modularity in terms of matrix values which define the optimization job in the form of linear algebra. It expressed that the detection of community structure can be possible by segregating the network into such communities that possess high modularity value. The eigenvalues and eigenvector of modularity matrix was well suitable for finding the modularity of a network. It showed a time complexity of O (n^2) for sparse networks.

2.3. Multilevel algorithm

Multilevel community algorithm aims to find the community structure in a large network. According [2], it was based on the optimization algorithm of modularity and is hierarchical in nature. It consisted of two phases- at the first phase, each community was being assigned to each node and modularity gain was calculated at each neighbor's end. The relocation of the node happened if modularity gain came out to be positive. Thus, all the nodes are shifted to some other community in a greedy approach. This phase stopped when the maximum value of modularity was obtained. In the second phase, a new graph was generated by re-assigning the nodes to the communities. The process gets terminated when the gain in modularity cannot be increased further. The multilevel algorithm has found to be advantageous in terms of computational time and complexity. It has shown easy implementation and fast and robust performance in large-scale networks. The InfoMap is a compression based technique that focused on Huffman coding [10] [3]. It represented the local interaction among the node and the random walk was used for the flow of information among the nodes. In this algorithm, time taken to visit each node by random walk was calculated that used a greedy approach followed by simulated annealing- a modularity optimization to clarify the results. The InfoMap algorithm has also shown the robust performance among the community detection algorithms.

Variance of Information (VI) : The variance of information is an entropy and information based on similarity measure was introduced. It computes the distance between the two partitions of a dataset chosen . It is defined as

$$VR(P/Q) = H(P/Q) + H(Q/P)$$

It exhibited some distance properties and acts as a touchstone for likening the partitions. [10]. The VI is an independent of the count of points in the dataset, so the comparison among the different datasets was easily achieved. The VI, being dependent on the count of clusters, grows largely as *2logK* where K is cluster count.

Normalized Mutual Information (NMI) : - It is defined as a clustering measure so that two different partitions of a dataset can be compared with each other to explore the common information among them. It generates the *confusion matrix* with a trade of real *v/s* found communities. It is defined as

$$IN(P, Q) = {}^{2\,IN(P,Q)}\!\big/_{H(P) + H(Q)}$$

Where IN is normalized mutual information (*nmi*) of two random variables, P and Q. If the two partitions are found to be identical, it generates an output 1, else 0. *Rand Index*- The Rand index was based on a count of pair of points. Given the two clustering A and B of a set of data points, say D, there can be four different categories to which these pairs of points. These categories are

a.M_{11}– the count of all the pairs that exists in the same cluster in both A and B,

b.M_{00} – the count of all the pairs that exists in the different cluster in both A and B,

Thus, the Rand index is defined as

$$\text{R and}\,(A, B) = \frac{M00 + M11}{N/2}$$

Where Rand (A, B) is a Rand index between the two clustering namely, A and B. So, the Rand index is defined as the ratio of all the nodes that are correctly identified in the two partitions with respect to the count of all the elements.

1. **Adjusted Rand index :** This is an extension of Rand index. It considers the expected index value and generates the output within a range of [-1, 1].
2. **Split join :** It is also split join distance between the different clusters.
3. **Modularity :** Modularity is a kind of quality measure that detects the quality of a partition and focuses on finding the community existence in a random graph.It is devised by Girvan and Newman [7] to find out the partition quality of the network. This is achieved by comparing the actual density of edges in a subgraph with the expected density without taking an account of the community structure. The null model being a replica of the original graph, helps to find out the expected edge density. The modularity can be defined as the summation of all pairs of nodes

$$M = 1/2e \sum_{mn} (A_{mn} - P_{mn})\, \delta\, (CR_m - CR_n)$$

Where M is the modularity measure, e is the total number of edges of a graph, A is the adjacency matrix, and P is expressed as the expected number of edges present between the two nodes ‘ *m*’ and ‘*n*’. The delta function δ generates an output 1 if two nodes ‘*m*’ and ‘*n*’ exists in same community CR, *i.e*., if $CR_m = CR_n$. The modularity measure varies with the size of the community.

3. EXPERIMENT AND RESULTS

This section describes the types of real-world datasets that were used for the empirical analysis. We chose social network dataset such that it covers all the major types of network. The dataset used here, consist of two Internet peer-to-peer network. In this first category, two graphs of collaboration networks, a network with ground truth communities, a social network and an email communication network. Overall, we considered 8 different datasets from several other sources. The first category of dataset chosen is Internet peer-to-peer network that consists of several Gnutella networks. The main purpose of Gnutella, being an open search protocol, is file sharing and signifies a kind of virtual network having some routing operations. The hosts in network topology are represented by the vertices (nodes) and edges represent the association between the hosts of the Gnutella network. Gnutella08 is a directed network having 6301 nodes and 20777 edges. The second category belongs to the online social network where several users create their accounts to share their information. The users join many communities according to their interest, hobbies and their professionalism, etc. The ego-Facebook network is an undirected network where the dataset consists of friend list of several people from Facebook. The social networking allows the users to generate their communal clique as per their interests and profession. The dataset has the information about the different profiles, ego-networks and their cliques. The Facebook network consists of 4039 nodes and 88234 edges with an adjacent clustering coefficient value of 0.6055. The ego-centric Facebook network. DBLP collaboration network with ground truth communities. It is a bibliography of computer science having plenty of research papers available It represents the authors as nodes and other co- authors being connected to them, are represented by edges. The authors having a same domain interest belongs to one community and the authors who have published their paper in the same conference generate a scientific community. So, the conference venue serves as a ground-truth community. It is a large undirected co-authorship network with 317080 nodes and 1049866 edges. YouTube network, a kind of social graph where the users can join as many groups as they want to. The average community size is 13.50 with 8.385 communities. It is a large undirected network with a clustering coefficient of 0.08.Email-Enron dataset is a communication network that consists of information about various email communication of a million users. The email address of the network represents the nodes and edges represent the email conversation sent by a node '*x*' to '*y*'. Another category of the network is ArXiv citation network that consists of GR-QC and HepTh network. GR-QC stands for General Relativity and Quantum Cosmology network, which considers the papers that are published in cosmology category, whereas the HepTh signifies High Energy Physics Theory collaboration network, which a collection of papers has published in physics theory category. Among the two datasets, GR-QC is the smallest network having 5855 authors as compared to HepTh having 9877 authors.

3.1. Plots of different graphs

The different types of networks analysis is done using R- programming language. it is evident that modularity is highest for the DBLP collaboration network when multilevel algorithm is applied and lowest for Email-Enron network when Infomap community detection algorithm runs over it. The modularity represents the quality of partition of the network so, the DBLP generates the best partition among the different real-world networks with multilevel algorithm. Considering the Walktrap algorithm for community detection, we found that the best split is generated in the DBLP collaborate on network, whereas the lowest quality of partition is seen on Gnutella Internet peer-to-peer network. The modularity value of DBLP is 0.81817 and Gnutella09 is 0.30901. When the eigenvector algorithm runs over the different real world network, the Ego-Facebook network shows the best result with a modularity value of 0.79913, whereas the partition quality is poor for DBLP network. The Ego- Facebook network outperforms best with Eigenvector algorithm and Infomap algorithm for detecting the communities in the network.

The multilevel algorithm for community detection generates the modularity value of 0.88348 for DBLP network and 0.46342 for a Gnutella Internet peer-to-peer network. After analyzing the modularity value of all the algorithms, we found that the multilevel community detection algorithm gives better quality of partition for all types of real-world network. It has been found that the best quality partition value is generated by DBLP network and lowest value is generated by Gnutella09 peer-2-peer network. The DBLP was poorer and Facebook engender the highest value of modularity with Eigenvector algorithm. The Multilevel algorithm generates the best modularity value for the DBLP collaboration network while the Infomap generates the best modularity value for CA-GrQc network. The above analysis leads to the fact that multilevel algorithm is suitable for all the networks of different sizes.

3.3. Communities Count table

In the following table 1, it depicted the total number of communities found in the real-world networks when different community detection algorithm runs over using R software. The study suggested that the largest number of communities is present in the DBLP collaboration network with Walktrap algorithm. The evidence from the analysis reveals that the number of communities depends on the size of the network and the algorithm used for finding the communities.

Here, EV signifies the Eigenvector algorithm, ML signifies then Multilevel algorithm, IM represents the InfoMap algorithm and WT is Walktrap algorithm. The Rand index varies with the count of clusters (communities) and the count of data elements. The count of communities increases and directly proportional to the Rand index value. The lowest number of communities using Eigenvector and Multilevel algorithms was observed in Internet peer-to-peer network and this count is increasing as we go downwards. The best performance is generated for a YouTube network with a value of 0.9989, 0.9969 and 0.9988 respectively by using the Rand index.

Table 1. Number of communities in different networks

Algorithm name	***Walktrap.***	***Leading.eigenvector.***	***Infomap.community***	***Multilevel.community***
Graph name	***community(WT)***	***com munity (EV)***	***(IM)***	***(ML)***
P2p-Gnutella08	0.33860	0.40745	0.35056	0.46382
P2p-Gnutella09	0.30901	0.40892	0.34780	0.46526
CA-GrQc	0.79332	0.78808	0.80702	0.86012
CA-HepTh	0.66370	0.60690	0.69705	0.76805
Ego-Facebook	0.81194	0.79913	0.76906	0.83478
Email- Enron	0.51179	0.43944	0.12945	0.60508

Considering the comparison of eigenvector with Walktrap, Imfomap and Multilevel algorithm, the Variance of Information (*vi*) has shown the best performance among all the four performance measures, ignoring split-joint measure, for a small sized network, *i.e.*, Gnutella peer- to-peer network. For small sized network, vi generates an output of a 4.637 value when an EV is compared with the ML algorithm, a 5.545 value when an EV is compared with the IM and 3.4507 value when EV is compared with the WT algorithm. Since, the Normalized mutual information is used to evaluate the comparison between the two modules of a real- world social network. Therefore, the evaluation of nmi is performed against four different algorithms, namely, Walktrap algorithm, Eigenvector algorithm, Multilevel algorithm and Infomap algorithm and observed that Normalized mutual information (NMI) has given the best performance for a YouTube real network with a value of 0.9927 against EV and ML algorithm, 0.9867 against EV and IM algorithm and 0.9909 against EV and WT algorithm. The paper[4] explained that the Variance of Information (*vi*) is defined as a performance measure that is local in nature. This implied that the similarity between the two clusters depends only on their difference value and not on the rest of the network. For small sized networks, vi generates a low value of similarity, *i.e.*, the ArXiv citation network has generated small value. The vi measure has generated a lowest value of 0.187 for a YouTube network when Eigenvector is compared with Multilevel algorithm, a value of 0.346 for a YouTube network when Eigenvector is compared with Infomap algorithm and a value of 0.2348 for a YouTube network when Eigenvector is compared with Walktrap algorithm. Also note that, the vi measure generates a larger value for the email communication network for Eigenvector,

Walktrap and Infomap algorithms except a case where a larger value is generated in Internet peer- to-peer network. This implied that as the size of network increases, the vi value decreases exhibiting the best performance for small sized networks. Adjusted rand index is an extension of Rand index. In paper[4] explained that the adjusted random index does not depend on the way of partitioning of the whole network. The adjusted rand index outperforms best in ego-Facebook social network. The ego-Facebook network has produced the value of 0.7876, 0.5401, 0.5922 values for EV and ML, EV and IM, and EV and WT algorithm respectively. It implied, the fact that the adjusted rand index varies with the size of the network and it increases as the size increases. Therefore, the Internet peer-to-peer network being smaller in size, shows the lower value of this measure and Facebook social network exhibits larger value because of large size. However, in the case of DBLP, the larger in size, but due to the division of structural groups shows week performance during analysis. It is an exception and does not lead to any conclusion.

The table 2 depicts that the large sized DBLP collaboration network and YouTube network has shown the best performance among all the networks. The DBLP generated the large value of 139077 when eigenvector and multilevel algorithm is taken into consideration for the comparison purpose and 148731 against the comparison of eigenvector with the walktrap algorithm.

Table 2. Experimental time calculation (in seconds)

NETWORK ↓	EV	ML	IM	Number of nodes	Number of edges
P2p- Gnutella08	7.91	1.45	93	6301	20777
P2p- Gnutella09	11.5	1.92	2.64	8114	26013
CA- GrQc	12.6	3.86	9.62	5242	14496
CA- Hepth	18.45	2.76	22.98	9877	25998
Email- Enron	20.76	2.71	55.36	36692	183831
Ego-Facebook	6.71	0.072	20.65	4039	88234
DBLP	41.48	49.60	7200	317080	1049866
YouTube	72	55	7500	1134890	2987642

Here, the table embodies the number of nodes and edges present in the network and the experimental time taken by each of the dataset to run the algorithm. The LV represents Eigenvector algorithm, ML represents Multilevel algorithm and IM signifies the Infomap algorithm. The experimental time calculation gives a light on the fact that Multilevel performs faster as compared to other algorithms and Infomap algorithm gradually becomes slower as the size of the network is escalated.

3.4. Plots depicting experimental time

In figure 1, the experimental time taken by each of the algorithms is presented. It represents the time with respect to the number of nodes present in the real network.

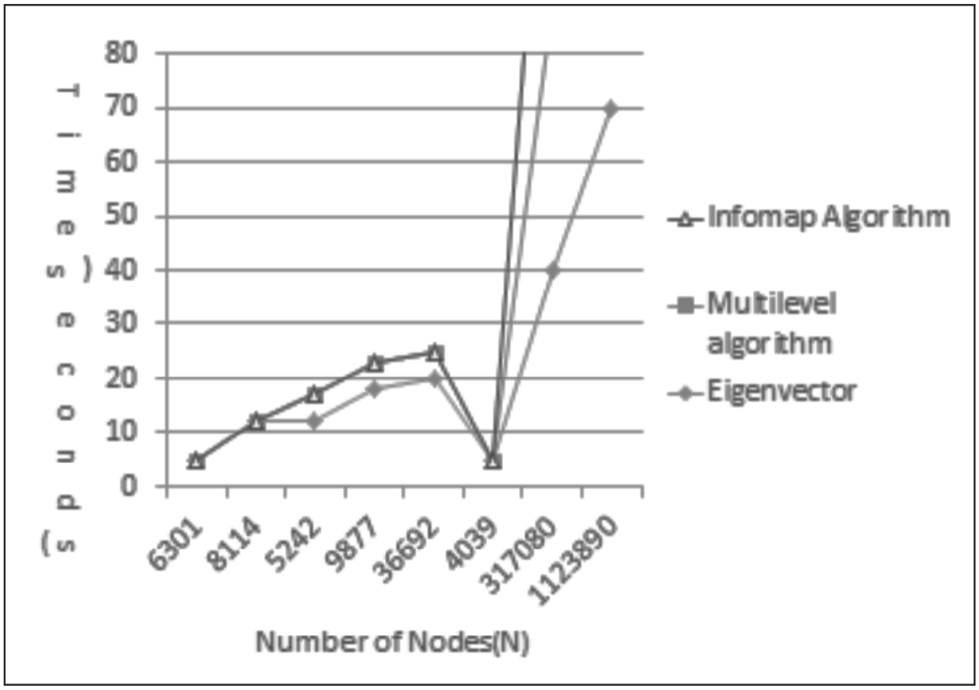

Fig. 1. Experimental time depiction of (*a*) Eigenvector algorithm (*b*) Multilevel algorithm and (*c*) Infomap algorithm

4. RESULTS AND DISCUSSION

The ever increasing size of the social network has enabled the researchers to focus on the analysis and detection of such communities that exhibited some desirable properties and Eigenvector algorithm, *b*) Multilevel algorithm, and *c*) Infomap algorithm. The multilevel algorithm runs in fractions of time and hence, showed the best performance than others. The large sized DBLP generates the best partition with Multilevel and Walktrap algorithm and the worst performance with eigenvector algorithm. It is evident from the above analysis that the multilevel algorithm is fast and robust for all types of networks and Infomap is quite time consuming. The Eigenvector and Infomap algorithm has given a high value of modularity for an ego-Facebook social network.

The community count table 1, depicted the number of communities present in real-world networks. The communities are counted against each community detection algorithm. It has been analyzed that the community count varies with the network size and the type of algorithm used for community detection. Therefore, the count of the community is largest in the large-sized DBLP network. The comparison table 2 has been generated for several networks with five different performance measures, namely, Normalized mutual information (*nmi*), Variance of Information (*vi*) , Rand index, Adjusted rand index and Split-join. It was observed that the vi measure has shown the best performance among all the measures excluding split-join and the YouTube network outperformed well using nmi, rand index and split-join performance measures. The DBLP collaboration network excels by using split-join but worst in the adjusted rand index. Since, the vi measure decreases with the increase in community size, the small sized networks manifested the best performance when several algorithms are compared for community detection. Thus, the multilevel algorithm outperformed well among all the algorithms used to find the community structure. The plots of real-world datasets are taken by using the R-programming language and the complete experimental analysis are performed on dissimilar sized real networks. The comprehensive description of all the networks has been provided in the paper. The experimental time taken by each of the networks to run the different algorithm has been shown in the table 2, which reveals the understanding of the networks not only in terms of modularity, but also in terms of the time as well.

5. CONCLUSION

The community structure in social network analytics plays an important role in unearthing the concealed information. So, it is essential to explore the community structure using the algorithms. In this paper, we have analyzed eight different real-world datasets using the four different community detection algorithms. The modularity value of each algorithm is obtained by using the R-programming language and we found that DBLP outperforms best with Walktrap algorithm and Multilevel algorithm. It also gives light on the number of communities found in dissimilar sized real networks and reveals the mere fact that the size of the networks helps to determine the community count. As the size of the network goes on increasing, the community count also gets increases. Consequently, the maximum number of communities is detected on the DBLP collaboration network, which possesses a large size. This analysis has been illustrated with the help of graphs/ networks. After an exhaustive analysis of the community count and the modularity value, the different algorithms are compared using the different performance measures and the value of each performance measure is calculated against each pair of algorithms. From this comparison, it has been found that the variance of information (vi) outperformed well among the other measures. The multilevel algorithm has shown the robust and efficient performance for all the networks. The comprehensive use of R-programming language leads to the inception of efficient values. The plots of this experimental time indicate the scalability and efficiency of the algorithms. In future, more data with more algorithms can be used to detect the performance of the

6. REFERENCES

1. Adar, Eytan, and Bernardo A. Huberman. “Free riding on gnutella.” *First Monday* 5.10 (2000).

2. Barabási, Albert-László, and Réka Albert. “Emergence of scaling in random networks.” science 286.5439 (1999): 509-512

3. Bohlin, Ludvig, et al. "Community Detection and Visualization of Networks with the Map Equation Framework." *Measuring Scholarly Impact*. Springer International Publishing, 2014. 3-34.

4. Fortunato, Santo. "Community detection in graphs." *Physics Reports* 486.3 (2010): 75-174.

5. Girvan, Michelle, and Mark EJ Newman. "Community structure in social and biological networks." *Proceedings of the National Academy of Sciences* 99.12 (2002): 7821-7826.

6. Massen, Claire P., and Jonathan PK Doye. "Identifying communities within energy landscapes." *Physical Review E* 71.4 (2005): 046101.

7. Newman, Mark EJ. "Finding community structure in networks using the eigenvectors of matrices." *Physical hidden review E* 74.3 (2006): 036104

8. Newman, Mark EJ, and Michelle Girvan. "Finding and evaluating community structure in networks." *Physical review E* 69.2 (2004): 026113.

9. Newman, Mark EJ, Steven H. Strogatz, and Duncan J. Watts. "Random graphs with arbitrary degree distributions and their applications." *Physical review E*64.2 (2001): 026118.

10. Orman, Günce Keziban, Vincent Labatut, and Hocine Cherifi. "Comparative evaluation of community detection algorithms: a topological approach." *Journal of Statistical Mechanics: Theory and Experiment* 2012.08 (2012): P08001.

11. Reichardt, Jörg, and Stefan Bornholdt. "Detecting fuzzy community structures in complex networks with a Potts model." *Physical Review Letters* 93.21 (2004): 218701.

12. Strogatz, Steven H. "Exploring complex networks." *Nature* 410.6825 (2001): 268-276.

13. URL snap.stanford.edu

14. Wagner, Silke, and Dorothea Wagner. *Comparing clusterings: an overview*. Karlsruhe: Universität Karlsruhe, Fakultät für Informatik, 2007.

15. Ward Jr, Joe H. "Hierarchical grouping to optimize an objective function."*Journal of the American statistical association* 58.301 (1963): 236-244.

16. Watts, Duncan J., and Steven H. Strogatz. "Collective dynamics of 'small-world'networks." *nature* 393.6684 (1998): 440-442.

17. Yang, Jaewon, and Jure Leskovec. "Defining and evaluating network communities based on ground-truth." *Knowledge and Information Systems* 42.1 (2015): 181-213.