International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 02 Issue: 04 | July-2015 www.irjet.net p-ISSN: 2395-0072

Studying the Properties of Complex Network Crawled Using MFC

Varnica[1], Mini Singh Ahuja[2]

[1] *M.Tech(CSE), Department of Computer Science and Engineering, GNDU Regional Campus, Gurdaspur, Punjab, India*

[2] *Assistant Professor, Department of Computer Science and Engineering, GNDU Regional Campus, Gurdaspur, Punjab, India*

Abstract - *Online Social Networks (OSNs) have become an important subject of research in almost all areas of sciences like computer science, social sciences etc. This presents a need for valid and useful datasets which can be used in research. The time taken to crawl the entire network introduces a bias which should be minimized. Usual ways of addressing this problem are sampling based on the nodes (users) ids in the network* ***or crawling the network until one "feels" a sufficient*** *amount of data has been obtained. In this paper, a survey on various crawling procedures has been done. From the survey, it has been found that none of the procedure gives efficient results. Therefore the paper ends with a future scope to overcome these issues. .*

Key Words: Online Social Networks, BFS, DFS and MFC etc…

1. ONLINE SOCIAL NETWORKS

Networks are present in each and every aspect of our lives. We are surrounded by a number of networks like WWW (World Wide Web) is a network which we use every day. The friendship between individuals, the business relations etc. are all networks.

Online social networks are very large and complex dynamic networks. They have grown at a very large rate in the last decade. Previously, the online social networks were limited to only a hundred of people but now, about a million of people are part of one network. Face book and twitter are examples of such networks. Also, in earlier researches online social networks were considered to be static because of which many features were ignored. Now, with a vast growth in the number of users in such networks, more emphasis is laid on the dynamic nature. These networks have many properties like associative mixing, community structure etc. which can be studied to extract useful information from these networks.

Analyzing such networks has become very important as they are of interest to many different areas like sociology, marketing, engineering, medical etc. The diverse use of online social networks makes it important for the network providers to understand how the traffic is generated with the different activities of the users. The graphs generated by such networks are very large and make it very difficult to completely understand it [17].

2. COMPLEX NETWORKS

Complex systems are basically networks which are organized into different compartments such that each compartment has its own role and function to perform. All the compartments consist of nodes. The links between nodes are of high density whereas the links between compartments are of low density [5].

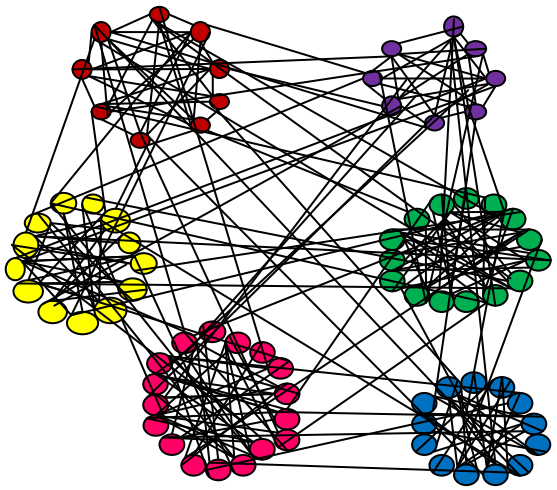

Fig -1: A Complex Network with Different Communities

A complex network represents the interactions in the real world in the form of a mathematical model. The major problem that occurs in complex networks is the identification of the communities which are hidden in the structure of these networks.

Extracting data from these networks and testing the community detection algorithms is very difficult. It is very costly and time consuming to obtain real world data. Moreover, the complex networks have many properties such as average degree, shortest path, degree distribution etc. which are very difficult to be controlled in real world networks [11].

Real world data can be generated by using a methodology called web crawler. Web crawlers are also known as web spiders.

The main aim of a community detection algorithm is to divide nodes or vertices of a network into any number of communities or groups, maximize the number of edges between groups and minimize the number of edges between vertices in different groups [16]. Till date many community detection algorithms have come up to detect communities.

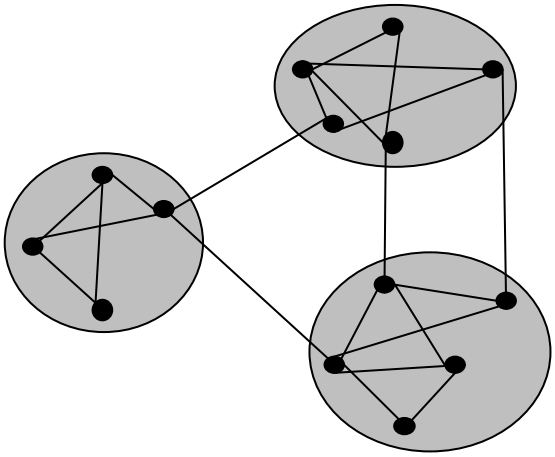

Fig -2: Links between Different Communities

3. PROPERTIES OF COMPLEX NETWORKS

Complex networks are very popular and is an emerging topic for research. These networks exhibit many properties which have been studied in the past. Some of these properties are:

3.1 Small World Effect

A complex network is said to exhibit the property of small world effect based on the condition that the average path length between the nodes in the network is less. This concept of small world effect has been originated from the observations of the experiments carried out by Milgrame in the year 1967. In his observations it was found that the letters which were passed from one person to the next, reached at the target in small six steps.

This property is also known as "The Six Degree of Separation Principle". The model proposed by Watts and Strongatz is the most famous model exhibiting small world effect and is named as Watts and Strongatz small world model [20]. Small world effect property can be seen in many real world networks like metabolic networks, network of road maps etc.

3.2 Scale Free Network

Barabasi and Albert introduced the concept of scale free networks. Scale free networks are those networks which have power law degree distribution. Examples of scale free networks are biological networks, social networks etc [19].

3.3 Assortativity

The assortativity coefficient 'r' is a measure of the connectivity among the nodes in a network which are common in some way like nodes with similar degree. Networks like social networks show assortativity as highly connected nodes tend to be connected to nodes with high degree whereas networks like biological networks show dissortativity as the nodes with high degree tend to be connected with nodes of low degree. The assortativity coefficient r varies from 1 to -1. If r=1, it means that that the nodes tend to be connected to nodes with similar degree but if r=-1, it means that the nodes tend to be connected to nodes with varying degree [19].

3.4 Community Structure

A community is a set of entities which are linked to all the other entities in the network. The entities in one community perform the same function and share some common properties. A community structure reveals the internal organization of the nodes. Different communities combine to form a complex network.

Networks like biological networks and social networks reveal modular structure [22]. These structures exhibit more connections within a community than between different communities. The model proposed by Girvan and Newman was the first model which generated networks with the property of community structure.

3.5 Clustering Coefficient

Clustering coefficient is defined as the ratio of number of directed links that exist between the neighbours of a node to the number of possible links that could exist between the neighbours of that node. The clustering coefficient of a network is the average clustering coefficient of all the nodes in the network [19]. For ith node of a network, clustering coefficient can be calculated as,

$$C = \frac{2e}{k(k-1)}$$

where, K is the number of neighbours of ith node and e is the number of edges between these neighbours. The value of clustering coefficient lies between 0 and 1. The higher value of clustering coefficient indicates that there is higher degree of "cliquishness" between the nodes of the network. The '0' value of clustering coefficient for a graph indicates that it has no "triangles" of connected nodes and if the value of clustering coefficient for a graph is '1' then it is a perfect clique.

3.6 Degree Distribution

A network consists of large number of nodes and all the nodes have varying degree. Degree distribution P(k) for a graph gives the probability of a randomly selected node to have a degree 'k' in the network. It is used to describe the distribution of the links among the nodes in the graph.

3.7 Density

The size of the network can be known from the total number of nodes in the network. Density is used to define the level of linkage between the nodes in a network. It is

generally calculated as the ratio of the number of existing links between the nodes to the number of possible links. Mathematically, it can be written as,

$$\text{Density} = \frac{2E}{N(N-1)}$$

where, E is the edges of the network and N is the number of nodes in the network. For a complete network i.e., for a network in which all the nodes are connected with each other, the value of density is 1 [19].

3.8 Node Centrality

To check the importance of nodes in a network different measures are available. Networks can be directed, undirected or weighted. Different types of centrality measures are:

- Betweeness Centrality: It is defined as the number of shortest paths running from a given node. This metric is used to check the importance of an edge such that edges with high betweeness centrality lie on shortest paths and are more important in concern with the structure of the graph.
- Closeness Centrality: This metric is used to check the importance of a node. It is defined as the inverse of the sum of the distance between a node and all other nodes. This node is considered important if it is relatively close to all other nodes in the graph.
- Degree centrality: It is defined as the number of edges of a node.
- Eigen Vector Centrality: Eigenvector centrality measures the influence of a node in a network. It allocates relative scores to all nodes in the network [19].

4. LITERATURE SURVEY

Lancichinetti A. et al. (2008) [1], introduces a class of networks that explains the heterogeneity in the distribution of node degrees and community sizes. They have conducted a test for modularity optimization and a clustering technique which is based on Potts model. The results show that the performance of the algorithm is greatly affected by the size of the graph and the density of the links.

Lancichinetti A. and Fortunato S. (2009) [2], further continued their study and tested their benchmark on directed and un-weighted graphs. They have also paid attention to overlapping communities which is an important characteristic of community structure in real world networks.

Gunce K. Orman and Labatut V. (2009) [4], have tested different community detection algorithms to generate a set of artificial networks with different sizes and properties, and then analyze the different algorithms. It includes some explanation about what are the different properties of a complex network. The results show that based on the information used by these authors, spinglass and label propagation algorithms show the best results.

Coscia M. et al. (2010) [7], organizes the different categories of community discovery methods based on the definition of community adopted by them. They have discussed several problems like impact of no universally accepted definition of community on community detection task, overlapping communities in real networks etc. and have tried to find solutions for these problems.

Olston C. et al. (2010) [6], presented the basics of web crawling. The crawling architecture is discussed in detail and also information about the future scope of crawling is provided by the author. They have also elaborated on how the undesirable content can be avoided and also discusses the future directions in this field.

Gunce K. et al. (2013) [16], has tried to overcome the drawback of LFR benchmark as this benchmark does not produce all the features of real world networks. With two modifications in the algorithm, the results show that the centralization and degree correlation values of the generated networks and the real world network are very close. Also, the detection of different communities becomes difficult as the proportion of inter-community links increases. It performs testing of several community detection algorithms with the modifications applied on LFR.

Khurana D. and Kumar S. (2012) [11], studied different reserches on web crawler. Different search engines and web crawling techniques have been discussed in detail, on the basis of which they have presented general web crawler architecture, robot exclusion principle and different data structure used for crawling purpose. Thay also give a brief information about the working of different crawling techniques being used by many search engines.

Blenn N. et al. (2012) [10], introduced a new way of crawling large online social networks. This technique was named mutual friend crawling. It is compared with standard methods of crawling using breadth first search and depth first search. This was the first analysis of crawling toward community structure. In this method, the communities can be analyzed by the researchers even when the crawling process is running. Future work is required in terms of existence of overlapping communities in the network.

Iswary R. et al. (2013) [14], discusses the different techniques to develop a crawler and how to build an efficient crawler. They also elaborate on different crawling techniques like focused crawler, distributed crawler, incremental crawler and hidden web crawler. Also, the different design issues have been discussed in detail. They have presented the architecture of focused crawling and incremental crawling. The authors concluded that the

incremental crawling gives better performance as compared to other crawling techniques in terms of revisiting the pages.

Yang J. and Leskovec J. (2013) [15], proposed the concept of ground truth communities which provides interesting future directions. They have studied the 13 different structural definitions of the network and performed tests on their sensitivity, robustness and performance for identifying ground truth communities. They provide a parameter-free community detection algorithm which can be used for a network with more than 100 million nodes.

Yan H. et al. (2013) [12], present the community detection algorithms using local and global information. The information is extracted based on local network structure using local similarity measure and global network structure using betweenness. They try to increase the difference between inter-community and intra-community edges so that a more clear community structure is available with us and is easily detectable. For this purpose, they use the concept of redefining the weights of the edges between the nodes. The testing of the proposed algorithm has been performed on artificial as well as real world networks. The proposed algorithm is applied on Girvan-Newman benchmark, LFR network and Zachary's karate club. The results obtained are then compared with the original Girvan-Newman algorithm. The results of local similarity index which are based on local random walk dynamics are better than those which are based on local cyclic structures.

5. WEB CRAWLING

In case of World Wide Web, the network generated depends on the web crawler used for sampling. Web crawlers generate the graph structure of the web. To a web crawler, the web seems to be seems to be a large graph with pages at its nodes and hyperlinks at its edges [6]. A crawler starts at a few of the nodes (seeds) and then follows the edges to reach other nodes. Frontier contains the URLs of unvisited pages and in terms of graph, it is a list of unvisited nodes [13].

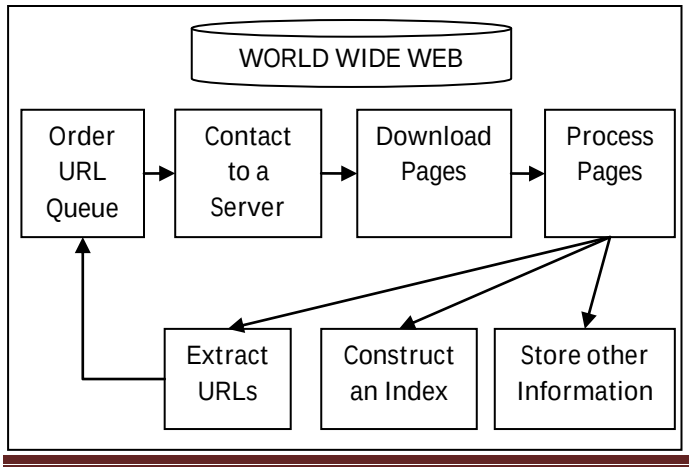

Fig -3: Basic Process of a Web Crawler

The figure above shows the basic process of a web crawler. Crawling starts with a collection of URL addresses and continues until it comes to a dead end or until some restriction defined in the crawling policy of the search engine is met. The crawler connects with the server to download documents. Words extracted from the documents are indexed and the URLs extracted are added to the URL queue and can be accessed whenever required [14].

Crawling is the most common approach of analyzing structures in online social networks like face book, to gather the network (by crawling) and afterwards partition the network into groups or communities by community detection algorithms. However, it is a very time consuming task to crawl the entire social network. Analyzing the network with community detection algorithms can be computationally expensive. Therefore, we design artificial networks [10].

6. WEB CRAWLING TECHNIQUES

6.1 Breadth First Search Crawling

This technique of crawling is used to find shortest path in unweighted graphs. The data structure used in this technique is queue. The general idea of this technique is that it begins with a starting node such as 'A' and then examine all the neighbours of 'A'. this means that it visits all the successors of a visited node before visiting any child of any of these successors. Crawling using BFS creates wide and short trees [14]. This is more commonly used than DFS.

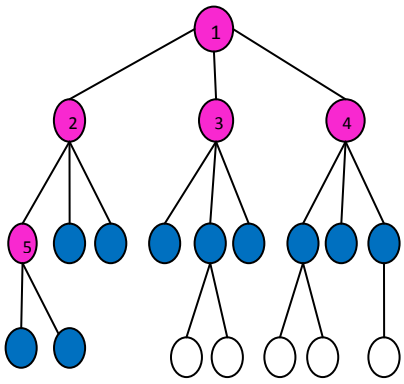

Fig -4: Breadth First Search

6.2 Depth First Search Crawling

This technique is used to verify if there is a path between two nodes. The data structure used in this technique is stack. The general idea of this technique is that it visits the successor of a visited node and before visiting any of it's brother node, it visits the child nodes of that successor. Crawling using DFS creates very long and narrow trees. This is less commonly used than BFS [14].

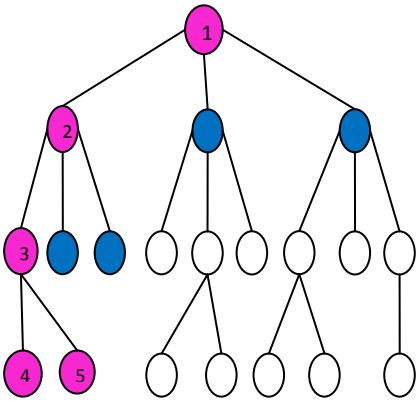

Fig -5: Depth First Search

6.3 Mutual Friend Crawling

Mutual friend crawling (MFC) was introduced by Blenn N. et al. . This approach crawls all the nodes of a network in such a way that all the communities are visited one after another. This algorithm assumes the knowledge of the degree of the neighbouring nodes which is a very difficult task in online social networks. The communities found using mutual friend crawling are smaller as compared to those found by other methods but the properties are same[9].

MFC is generally based on breadth first search algorithm with two differences. The first difference is that a map is used to store all the discovered nodes. The second difference is in the way the next node is chosen to crawl the network. For this, a reference score is calculated[10]. Reference score of a node is given by the ratio of number of discovered nodes to all the nodes which are linked to it. The list of all the discovered nodes is prepared and the one having highest value of reference score is being processed next [10].

The Figure 6 shows that an entire community is crawled before crawling other connected communities. Nodes are labelled based on the order of traversal during the crawling process. Different colours of the nodes denote different communities.

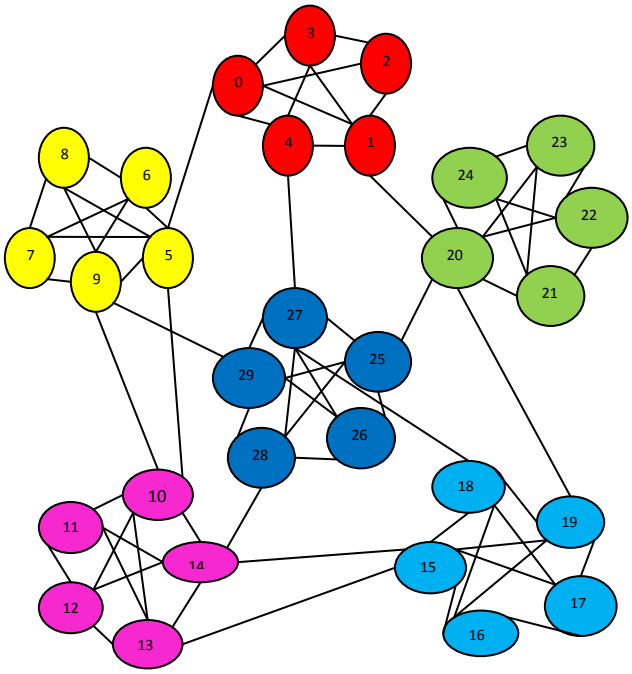

Fig -6: Mutual Friend Crawling

7. EXPERIMENTAL RESULTS

We have implemented mutual friend crawling in MATLAB using the Girvan and Newman's "American College Football Games" dataset. The number of clusters found in our result are 3. The following results show the different values of some properties of complex network like average path length, average clustering coefficient and average degree distribution.

Table -1: Computing values of different properties of complex network on a dataset using MFC

Average Path Length (APL)	Average Clustering Coefficient (ACC)	Average Degree Distribution (ADD)	Execution Time
1	1	14	0.4806
1.2857	0.68889	10	0.8134
1.7143	0.6	6	1.1225
1.7143	0.6	6	1.4433
2.2857	0.5	4	1.7936

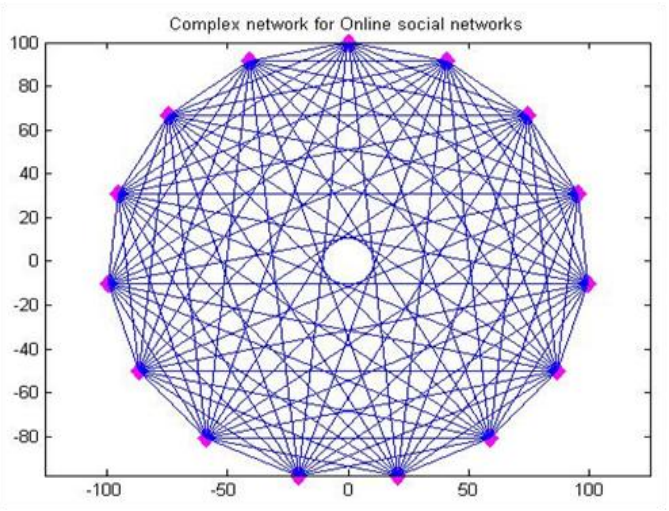

Fig -7: Complex Network with APL=1, ACC=1 and ADD=14

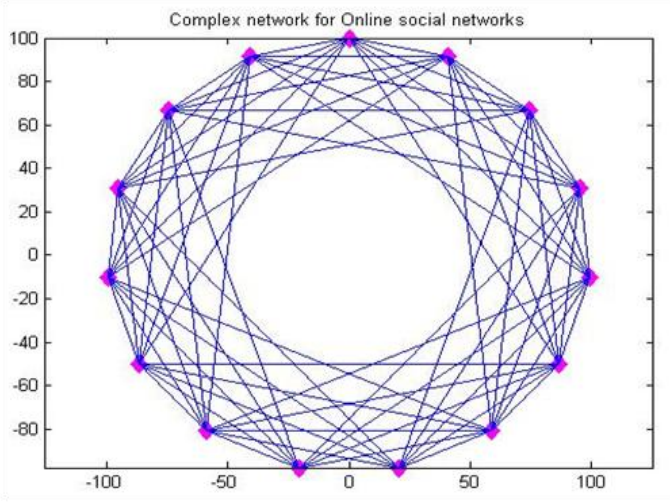

Fig -8: Complex Network with APL=1.2857, ACC=0.68889 and ADD=10

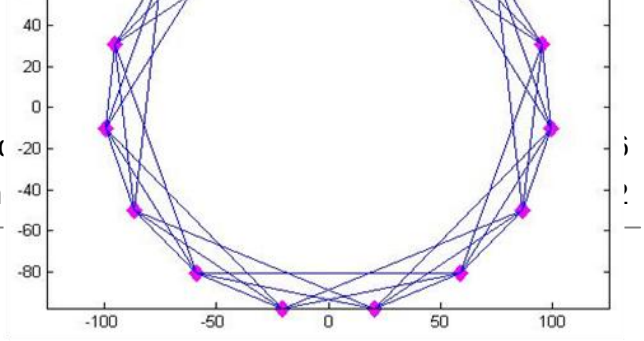

Fig -9: Complex Network with APL=1.7143, ACC=0.6 and ADD=6

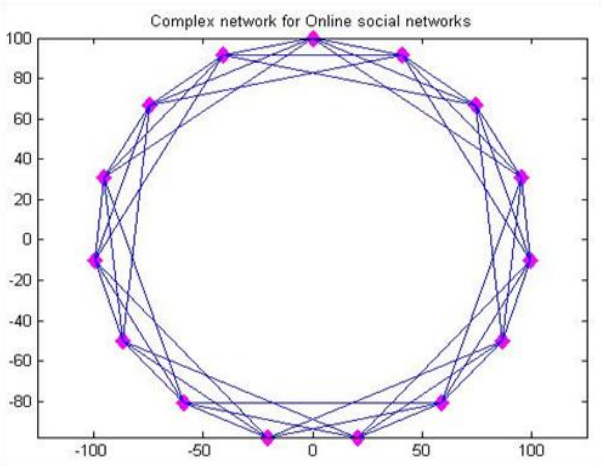

Fig -10: Complex Network with APL=1.7143, ACC=0.6 and ADD=6

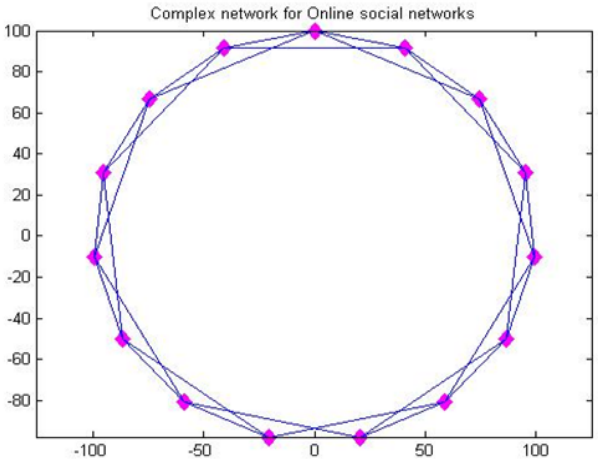

Fig -11: Complex Network with APL=2.2857, ACC=0.5 and ADD=4

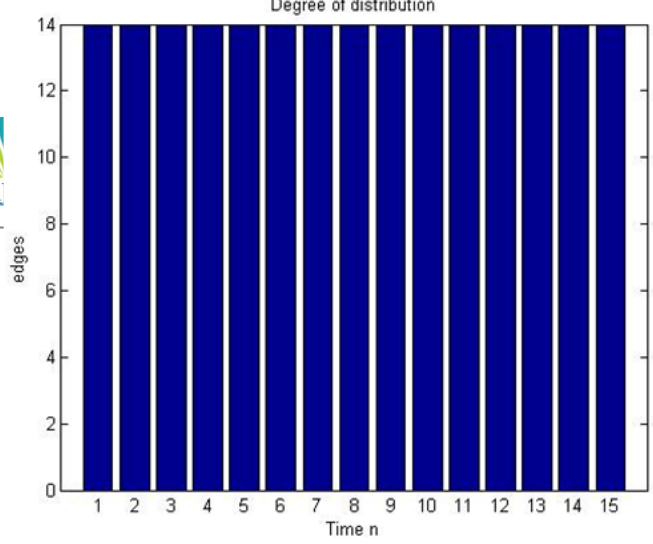

Fig -12: Graph showing degree of distribution

In our results, the complex network with different values of APL, ACC and ADD is shown in figures 7, 8, 9, 10 and 11. With the changing values the density of the network also changes. The nodes are represented with pink colour showing different teams in the dataset and the links between them is represented with blue colour edges. Figure 7 is the most dense network. Figure 12 shows the graph of average degree distribution corresponding to the network in figure 7.

8. GAPS IN LITERATURE

1. The most of existing crawling techniques are based on bivalent logic, the use of multivalent logic has been ignored.
2. The use of fuzzy based communities has been neglected in the majority of research.

9. CONCLUSION AND FUTURE SCOPE

In this paper, a survey on various crawling techniques has been done. From the survey, we have concluded that most of existing crawling techniques are based on bivalent logic and the use of multivalent logic has been ignored. Moreover the use of fuzzy based communities has been neglected in the majority of research. Therefore in near future, a fuzzy based clustering technique can be designed to obtain best communities.

REFERENCES

[1] Lancichinetti A. et al., *"Benchmark graphs for testing community detection algorithms"*, Physical Review E 78, 046110 9 (2008).

[2] Lancichinetti A. and Fortunato S., *"Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities"*, Physical Review E80, 016118 (2009).

[3] Lancichinetti A. and Fortunato S., *"Community detection algorithms: A comparative analysis"*, Physical Review E 80, 056117 (2009).

[4] Orman G. and Labatut V., *"A Comparison of Community Detection Algorithms on Artificial Networks"*, Discovery Science, Porto : Portugal, DOI : 10.1007/978-3-642-04747-3_20 (2009).

[5] Wang Z., *"Community Detection Approaches In Complex Networks: A Review"*, Department of Applied Mathematics, Fudan University.

[6] Olston C. and Najork M., *"Web Crawling"*, now the essence of knowledge, Vol. 4, No. 3 (2010) 175–246 (2010).

[7] Coscia M. et al., *"A Classification for Community Discovery Methods in Complex Networks"*, Wiley Online Library, Volume 4, DOI:10.1002 (2010).

[8] Orman G. et al., *"Qualitative Comparison of Community Detection Algorithms"*, International DICTAP 2011, DOI 10.1007/978-3-642-22027-2_23 (2011).

[9] Van Kester S., *"Efficient crawling of community structures in online social networks"*, PVM 2011-071, Tu Delft (2011).

[10] Blenn N. et al., *"Crawling and Detecting Community Structure in Online Social Networks using Local Information"*, Springer, LNCS 7289, pp. 56-67 (2012).

[11] Khurana D. and Kumar S., *"Web Crawler: A Review"*, International Journal of Computer Science & Management Studies, Vol. 12, Issue 01, ISSN: 2231 – 5268 (2012).

[12] Yan H. et al., *"Community detection using global and local structural information"*, Pramana – J. Phys., Vol. 80, No. 1 (2013).

[13] Pant G., *"Crawling the Web"*, The University of Iowa, Iowa City IA 52242, USA.

[14] Iswary R. and Nath K., *"Web Crawler"*, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 10, ISSN: 2278-1021 (2013).

[15] Yang J. and Leskovec J., *"Defining and Evaluating Network Communities Based on Ground-Truth"*, Springer, DOI 10.1007/s10115-013-0693-z (2013).

[16] Orman G. et al., *"Towards Realistic Artificial Benchmark for Community Detection Algorithms"*, International Journal of Web Based Communities, pp. 349-370 (2013).

[17] Gjoka M. et al., *"Practical Recommendations on Crawling Online Social Networks"*, IEEE Journal on Selected Areas in Communications, Vol. 29, No. 9 (2011).

[18] Varnica et al., *"Web Crawler: Extracting the Web Data"*, International Journal of Computer Trends and Technology, Volume 13, number 3, ISSN: 2231-2803(2014).

[19] Mini Singh A. et al., *"Future Prospects in Community Detection"*, International Journal of Computer Science Engineering and Information Technology Research, Vol. 4, Issue 5, ISSN: 2249-6831 (2014).

[20] M.E.J. Newman, *"The Structure and Function of Complex Networks"*, SIAM Review 45, 167-256 (2003).

[21] S. Fortunato, *"Community Detection in Graphs"*, Physics Reports, 486(3-5):75 (2010).

