# Community Detection Algorithms for Big Data using Graph Theory

**Ram Milan, Diwakar Shukla, Kamlesh Kumar Pandey**

***Abstract*: *Community detection is a nowadays research problem in the Big Data era related to huge volume, variety, and velocity of data. Big data defines data where normal processing, storage, retrieval fails and require some advanced tools to solve these types of problem. An important tool in the analysis of complex network is community detection. Community detection or community mining is a technique which is used to find the same type of relations in a particular group. Community detection is also known as Graph Clustering. This paper represents Big data in the form of graphs and detects community via some graph algorithms like METIS, Spectral Partitioning, hierarchical clustering, Markov Clustering, Genetic Algorithm based community detection algorithm, etc. Community detection is widely used in various types of disease detection, drug formation, species clustering. It can be also used in social networking sites to control crimes by detecting community bad peoples.***

***Keywords: Community Detection, Big Data, Graph Clustering, Markov Clustering***

## I. INTRODUCTION

Community detection(mining) is a technique that is used to find the same type of groups or clusters within the large networks or big data. We can find the same type of people or community in the social networking sites as in Face book, Twitter, etc. [1] presents a generic overview of graph clustering, which can be considered as equivalent to community detection. Some Community detection algorithms are METIS, Spectral Clustering, Hierarchical Clustering, Information theory-based algorithm, Markov Clustering, Genetic Algorithm based Community Detection algorithm. METIS is a set of serial programs for partitioning graphs, partitioning finite element meshes, and producing fill reducing orderings for sparse matrices. The algorithms implemented in METIS are based on the multilevel recursive-bisection, multilevel *k*-way, and multi-constraint partitioning schemes are already proposed by researchers. Dongen et al. [2].proposed MCL algorithm is short for the Markov Cluster Algorithm, a fast and scalable unsupervised cluster algorithm for graphs (also known as *networks*) based on simulation of (stochastic) flow in graphs. Spectral partitioning has become one of the most successful heuristics for partitioning graphs and matrices. It is used in many scientific numerical applications, such as mapping finite elements calculations on parallel machines [3].

A today societies are living in a digital era where everything is in the form of digital information like IOT, information produced from sensors, mobiles, and social networking sites like Facebook, Twitter Instagram and many more and there are a large number of digital users and each user has more than one mobile with the 4G internet connectivity and users are sending messages and likes as Face book page very frequently or do comment. Each and every information is stored on the cloud and every second large data is generated and this data is beyond the capability of the traditional data processing capability so we termed this type of data as the big data. Big data contains structure, semi-structured and unstructured data. Dobre and Xhafa (2014) report that every day the world produces around 2.5 quintillions bytes of data [4]. A graph G=(V,E) is a collection of V={v1,v2,v3,v4} called vertices, and another set E= {e1,e2,e3,e4} are called edges. There are two types of graphs directed graphs and undirected graphs [5].

## II. RELATED WORK

There are two classes of community mining on graphs. First is static community detection and the second one is community mining on dynamic graphs [6]. Girvan and Newman gave a significant definition called Network Modularity which is used for quality metric for measuring the partitioning of a network into communities [7]. Each of us may participate in many social cycles according to humans hobbies, educational background, working environment, political leaders, athletes, actors, scientist, and family relationship. As a result, when the network is large and the overlapping is significant most of the existing algorithms, in general, will have high computational cost due to their heuristics optimization strategies [7]. The community is often regarded as kind of cohesive substructures such are cliques, n-cliques, n-clans, n-plexes as well as the quasi-cliques. Community detection is based on the enumeration of all maximal cliques. Each group of the overlapping maximal cliques is regarded as a certain clustering kernel.

Big data is characterized by 7V's as Volume, Velocity, Variety, Veracity, Variability, Value, Visualization. Volume means a large set of data that consists of data in the number of terabytes, petabytes, and zettabytes of data even more. Velocity means a large amount of data within a defined amount of time. For example Wal-Mart process million of transactions each hour

**Revised Manuscript Received on August 22, 2019.**

**Mr. Ram Milan,** Computer Science and Applications, Dr. Harisiingh Gour Vishwavidyalaya, Sagar, India.

**Prof. Diwakar Shukla,** Computer Science and Applications, Dr. Harisiingh Gour Vishwavidyalaya, Sagar, India.

**Mr. Kamlesh Kumar Pandey,** Computer Science and Applications, Dr. Harisiingh Gour Vishwavidyalaya, Sagar, India.

[8].(Cukier, 2010).Variety means data comes in a various format like structured, semi-structured, and unstructured and it is in the forms of audio, video, text, images, sound, pictures and many more[9]. Veracity feature is used to measure the accuracy of data and this data is used for the analysis of data to find the usage patterns. Variability means that data whose meaning is constantly changing. Variability feature is also used in sentiment analysis where the same tweets have a different meaning to the world and every person takes it according to their view. Sentiment analysis is a new research topic. Visualization means a data must be represented in the format that is easily readable and it is possible to analyze it very easily. For example, Wal-mart has millions of customers and every hour millions of transactions are taking place visualization means the data represented in the form of the tableau. Value means extracting knowledge or value from a large amount of data that can be structured, semi-structured and unstructured without a loss for end users. Figure 1 shows the 7 V's of big data [10].

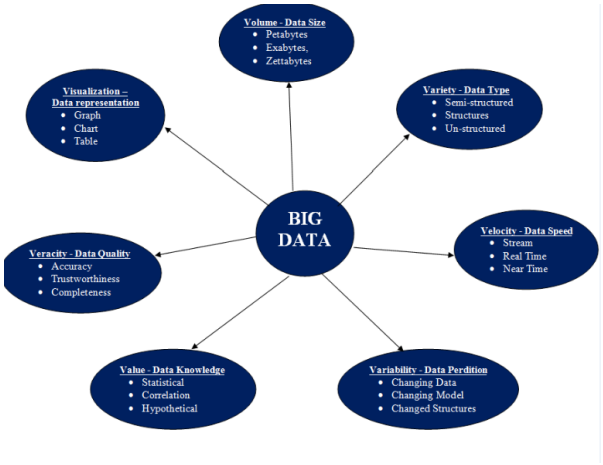


Figure 1 7 V's of Big Data

Big data as defined above have 7 V's, big data means at lest 3 V's are present then it can be treated as big data. Examples of big data Face book, Twitter, etc social networking sites where a large amount of data is generated every second that is velocity and the data generation speed is also very fast that is velocity and the data is also of very range like image, picture, files, music, video, etc that is variability. So Face book is a Big data example and in the Face book can be model as a graph and in the graph, we can also be used to find the communities. One community is also linked to another community we can also find this.

***2.1 Criteria for good community detection:*** there are three methods for finding good community detection techniques:-

***2.1.1. Modularity*** is firstly given by Newman [11]. That gives the portion of edges within community edges subtract the expected value of the same quantity for the randomized network. The formula of Modularity measure is $Q = \sum(e_{ii} - a_i^2)$ a fraction of edges from group iTo group j. $a^i$ is fraction an on of edges from/ to group I if the group by chance. If the number of within community edges is no better than random, we will get Q=0. Values approaching Q=1, indicating networks with strong community structure.

***2.1.2 Multi-criterion Scores:*** Jure Leskovec et al. [13] explains in the metrics that describe the notion of a quality of the cluster, Conductance, Expansion, Internal density, Cut Ratio, Normalized cut, Maximum-ODF, Average-ODF, Flake-of. Multi-criterion scores evaluate the communities from many aspects, which help to have a better understanding of the communities. Less value of score f(s)describes a more community as a set of nodes.

**2.2 Algorithms of Community Detection:** there are many methods have been proposed to find the community in a given graph. These methods are called as traditional methods of community detection, such as Clustering method, Newman's algorithm and so on. However, in real life, many people are belonging in one community and they are also related to other community as well so the many researchers are interested to find the overlapping communities detection. With the advancement of social networking sites and the exponential increase of the users the graph becomes more and more complicated and huge. Previous methods are focused on dividing the whole graph into a number of groups. So some researchers interested in study the local community detection algorithm [14].

***2.2.1 Traditional Algorithms of Community Detection:*** Researchers started to research as community detection in early 1970[15] and many researchers proposed a large number of algorithms and we are discussed some of the traditional algorithms of community detection.

***2.2.1.1.Partitional clustering:***Partitional clustering is a method of community detection. The algorithms assume there are k clusters in the social network and the aim is to find the points in the k clusters for to optimize a given cost metric between points or from points to centroids [16]. Some of the well-used functions are the k-clustering sum, k-center, k median, and Minimum k-clustering and so on.

***2.2.1.2 Hierarchical clustering:***Hierarchical clustering is divided into two types Agglomerative algorithms and Divisive algorithms [17]. The Agglomerative clustering works on clusters are iteratively merged bottom up if their similarity is sufficiently high, and the typical agglomerative algorithm is betweenness clustering. The divisive algorithm is that clusters are iteratively split top down by removing edges connecting vertices with low similarity.

***2.2.1.3 Newman's algorithm:*** Girvan and Newman proposed a series of classical methods [18]. Girvan et al. describe the method of edge betweenness that is the number of shortest route in whole graph under all vertex pairs with edges . The idea of these algorithms is to remove edges with the maximum betweenness measure and then recompute betweenness for all edges affected by the deletion.

***2.2.1.4 Graph partition:*** Graph partition divide nodes in a graph into a plurality of predetermined size communities which satisfies some objective functions by removing edges. Benchmark is the technique that is used to measure community structure. The planted l-partition

model is the easiest recipe. In this model one "plants" a partition consisting of a certain number of groups of nodes.

***2.2.1.5 Infomap:*** Infomap [19] was proposed to encampass the multipartite organization of large scale biological and social systems. The idea of this method is to use the probability technique flow of random walks on a network as a proxy for information flows in the real system and decompose the network into modules by compressing a description of the probability flow.

***2.2.2 Algorithms of overlapping Community Detection:*** Traditional algorithms are focused on identifying disjoint communities. However, in real life, people in the social network may belong to multiple communities [20]. So many researchers tend to research the overlapping communities detection algorithms and they proposed many algorithms.

***2.2.2.1 Clique percolation:*** The clique percolation methodology (CPM) [21] is that maybe first methodology to seek out overlapping communities. The concept of CPM is relating to communities as a collection of the maby cliques ( fully connected graphs). It begins by finding all k cliques.

***2.2.2.2 Link algorithm:*** Link algorithm is a link partitioning algorithm based on hierarchical partitioning clustering [22]. The idea of this algorithmic program is that two links share constant node to completely different communities, the node should be a node within the overlapping spce.Treat every edge as a seprate link community so merge the two most similar link communities till all the link communities become one community.

***2.2.2.3 COPRA*** **:** COPRA [23] could be an improvement of ancient label propagation rule and it is a multi label propagation rule. Label every vertex y with a group of pairs (a,c) wherever a could be a community symbol and c could ba a belonging constant. And so let the label unfold on the network supported the native structure of the network. Since every node is initialez with n labels, when propagation every node might have a multiple labels at constant time. Therefore Copra will determine overlapping communities.

***2.2.3 Algorithms of Local Community Detection:*** With the advancement of time and the advancement of the mobile devices a large number of users are using the social networking sites so the graphs become more and more complicated and huge so finding the communities also cost more. So some of the researchers proposed local community detection algorithms.

***2.2.3.1 Clauset's Algorithm:*** Clauset is that the initial to propose the matter of community detection. This algorithm is projected a definition of local measure. This algorithmic measure is easy and economical however must set community size in advance.

***2.2.3.2 Label Propagation Algorithm :***Label Propagation algorithm (LPA) [25] may be a dynamic methodology of community detection and this algorithm program relies on the native structure of networks to identify communities. Give every vertex a unique label and update each vertex x's label by replacing it by the label used by the greatest number of neighbors until the same label tends to become associated with all members of a community. The method has a low item complexity and it is easy to operate, but the algorithm has great uncertainty.

***2.2.3.3 Local Node Expansion:*** Local node expansion is to start with a number of nodes, and then according to the specific criteria expand the nodes to get the community. One of the most popular local node expansion is seed set expansion. Joyce JiyoungWhang et al. use the personalized Page Rank for seed expansion [26]. The algorithm computes Page rank scores, localized on seeds and then finds a set of high-rank nodes to form the community with the seed set. Another kind of method is based on node centrality.

***2.2.3.4 Local Optimization***

OSLOM [27] method optimizes the local statistical significance of communities. OSLOM consists of 3 phases. first, it looks for significant clusters, until convergence. Second, it analyzes the resulting set of clusters, trying to detect their internal structure or possible union. third, it detects the hierarchical structure of the clusters. OSLOM is the first method which is used in network accounting to detect clusters for edge directions, weights, overlapping communities and community dynamics.

## III. A COMPREHENSIVE STUDY OF THE COMMUNITY DETECTION ALGORITHMS

In this section this paper compares various community detection algorithms their merits, demerits, complexity, and scale through table 1. Tables gives 2 gives the clustering based community detection methods their author or algorithmic technique, method, and parameters. Like Newman and Girvan, Clauset et al, Blondel et al, etc.

Genetic algorithm uses heuristic algorithms to find the best solution for a given problem. These algorithms works on the principle on range of solutions known as chromosomes and fitness function is computed for these chromosomes. If it finds the optimum fitness one stops, else with some probability crossover and mutation operators are applied to the current set of solutions to find the new set of solutions. Community detection can also be used Genetic algorithm as an problem of optimization in which an objective function that captures the intuition of a community is a problem of optimization under the choosen better internal connectivity as compared to external connectivity. The table 3 is listed all the community detection algorithms based on Genetic algorithms.

| Algorithm | | Merits | Demerits | Complexity | Scale |
|---|---|---|---|---|---|
| Clustering | k-means | 1. Easy to implement<br>2. Average performance | 1. We need a number of clusters in advance<br>2. Terminates at a local optimum | $O(n^{dk+1})$ [28] | L |
| | Hierarchical Clustering | 1. There is no need to specify the number of clusters in advance. | 1. Where to cut the dendrogram tree we don't know.<br>2. If merging/division heuristics is not good we may get bad results. | $O(n^2)$ | M |
| Newman's Algorithm | G-N algorithm | 2. There is no need to specify the number of clusters in advance | 1. Where to cut the dendrogram tree we don't know.<br>2. This also is slow | $O(E^{2N})$ [29]. | S |
| | Newman's fast algorithm | 1. It is faster than the G-N algorithm<br>2. There is no need to specify the number of clusters in advance.<br>3. May get good partitions | 1. There is not a theoretical guarantee compared to the greedy algorithm | $O(n^2)$ [30]. | S |
| Graph Partition | Benchmark | 1. Computation is very less<br>2. Easy | 1. Difficult to agree on the same network | -- | L |
| | Spectral Clustering | 1. It gives very good results<br>2. It is very effective to handle complex shapes | 1. It is mostly not efficient<br>2. Not sure which objective is the right one to use | $O(n^3)$ [31]. | M |
| | Kernighan-Lin | 1. It is very fast | 1. Know the size of clusters in advance | --- | |
| Infomap | | 1. It uses information about weight and direction | 1. It considers only structural characteristics | O(E) [32]. | L |
| Clique Percolation | | 1. It can detect overlapping community | 1. It is used for networks with many full connected subgraphs | O(dm) [33]. | L |
| LINK Algorithm | | 1. It can detect overlapping community | 1. Don't know where to cut the dendrogram tree | --- | |
| COPRA | | 1. It can detect overlapping community | 1. it has great uncertainity | O(vm log (vm/n)) [34]. | L |
| Clauset's Algorithm | | 1. Efficient 2. Fast | 1. Know the size of clusters in advance | | M |
| Label Propagation Algorithm | | 1. It is efficient 2. Low time complexity 3. It can detect overlapping community | 1. It can detect one community | O(N) [35]. | L |
| Local Node Expansion | | 1. It has high accuracy<br>2. It is efficient<br>3. Niche targeting | 1. It considers only structural characteristics | -- | |
| Local Optimization | | 1. It can detect overlapping community<br>2. Can be generalized to directed graphs, dynamic networks, and weighted graphs | 1. May return slightly less accurate results than other methods | $O(n^2)$ [36]. | M |

Table 1.Comparisions of all the algorithms

| Author | Methodology | Parameters |
|---|---|---|
| Newman and Girvan[11]. | Divisive Clustering | Edge betweenness |
| Girvan and Newman[37][38][39]. | Modularity maximization | Modularity[37],eigenvector[38] and eigenvalue[39]. |
| Clauset et al [40]. | Greedy optimization of modularity | Edges, vertices modularity |
| Blondel et al (Louvain Method)[41]. | Hierarchical Clustering | Nodes,edges,Modularity |

| | | |
|---|---|---|
| Guimera et al [42], Zhou et al[43]. | Modularity optimization using Simulated Annealing | No. of links, linking probability, no. of modules, no. of partitions, Modularity [42] . No. of edges, inter factor and intra factor, Modularity[43]. |
| Dutch et al.[44] | Modularity optimization using External optimization | No. of nodes, links, degree, Modularity |
| Ye et al ( AdClust) [45] | Agglomerative clustering | Vertices, Force, Modularity |
| Wahl and Sheppard [46] | Hierarchical Fuzzy Spectral clustering | Fuzzy modularity, Jaccard Similarity |
| Falkowski et al (DENGRAPH)[47] | Density-based clustering | Distance function |
| Dongen et al [2] | Markovian clustering | Number of nodes |
| Nikolaev et al [48] | Entropy centrality based clustering | Transition probability matrix for the Markov process |
| Steinhauser et al [49] | Consensus clustering | Random walks similarity matrix |

Table 2: Clustering based on community detection methods.

| Author(Algorithm) | Approach | Parameters |
|---|---|---|
| Pizzuti(GA-Net) [50] | Community score as the fitness function | Community score |
| Pizzuti(MOGA-Net) [51] | Multi objective optimization. | Community score community function |
| Hafez et al [52] | Single objective, multi-objective optimization | Number of genes, mutation Crossover operators |
| Mazur et al [53] | Community score and Modularity as fitness functions | Fitness functions |
| Liu et al [54] | Genetic algorithm and clustering | Size of population, maximal generation number, maximum no. of generations for unimproved fittest chromosome fraction of mined hubs, no. of communities |
| Tasgin et al [55] | Modularity optimization | Modularity, population size, number of chromosomes |
| Zadeh[56] | Multi population cultural algorithm | BS-average, BSN |

Table 3: community detection algorithms based on Genetic algorithms.

Label propagation in a network is the propagation of a label to various nodes existing in the network. Each node attains the label possessed by a maximum number of the neighboring nodes. Table 4 discusses some label propagation based algorithms for discovering communities, their approach, application, and parameters.

Clique-based methods for detecting overlapping communities are discussed in table 5. Table 5 discusses algorithm like Palla et al, Lancichinetti et al, Du et al, Evans et al, Lee et al, Gregory et al, etc. table uses Clique percolation method for discovering communities.

Nonclique methods to discover overlapping communities are discussed in table 6. Table 6 discusses algorithms like Nicosia et al, Lancichinetti et al, Baumes et al, Chen et al, Alvari et al, Shi et al, etc.

| Author | Methodology | Application/Improvements | Parameters |
|---|---|---|---|
| Raghavan et al [57] | Iterative label propagation | SLPA[58], WLPA[59], COPRA[60], Label Rank[61], BMPLA[62] | Nodes,label [58], Labels,threshold [59], Label,similarity [60], nodes [61] |
| Xie et al (Label Rank) [63] | Propagation, inflation, cut off, and conditional update. | LaebelRankT | Belongingness coefficient, the threshold nodes [63]. |
| Wu et al (BMPLA)[64] | Label propagation, and overlapping communities | ---------- | Number of vertices, label to which vertices belong, the average degree |

Table 4 Label propagation based community detection

| Author | Methodology | Parameters |
|---|---|---|
| Palla et al (CPM)[65] | CPM | Nodes, threshold weight |
| Lancichinetti et al [66] | Works on the principle of the Fitness function | Fitness function |
| Du et al. (Com Tector)[67] | Kernels based clustering | Set of all kernels |
| Shen et al(EAGLE)[68] | Agglomerative hierarchical clustering | The similarity between the two communities |
| Evans et al. [69] | Line graph, clique graph | Edges,Partition |
| Lee et al. [70] | Cliques based expansion | Fitness funciton |
| Gregory et al. (CONGO[71], Peacock algorithm [72] | Split between | Local betweenness, short paths [71], the ratio of max, edge betweenness, and max,splitbetweeness [72]. |

Table 5.clique based methods for overlapping community detection

| Author | Methodology | Parameters |
|---|---|---|
| Nicosia et al.[73] | Network Modularity algorithm for overlapping communities | Degree of nodes |
| Lancichinetti et al(OSLOM)[74] | Edge direction, weights, hierarchy | N vertices, E edges, degree of subgraph, internal and external degree of subgraph |
| Baumes et al [75] | Clusters of overlapping vertices | Internal edge intensity, external edge intensity,internal edge probability,edge ratio,intensity ratio |
| Chen et al[76] | Game theory | Number of communities gain function, loss function |
| Alvaro et al[77] | Game theory based | Number of snapshots, with V vertices and E edges |
| Shi et al[78] | Objective function: partition function | Size of population,runninggenereation fraction of crossover,fraction of mutation |
| Xing et al (OCDLCE)[79] | Community detection | Nodes,edges,neighbours of node |
| Bhat et al ( OCMier) [80] | Density-based | Threshold |

Table 6 : Nonclique methods to discover overlapping communities

## IV. General Community Detection Problems arises in Big Data

Some traditional methods are easy but often need to decide the number or the size of clusters, other traditional methods are usually slow or need to know where to cut the dendrogram tree; Overlapping community detection methods can detect overlapping community but some methods are suitable for networks with many full connected subgraph or great uncertainty. Local methods are efficient or with low time complexity but there are exist some shortcoming such as can detect only one community or need to know the size of-of the clusters in advance. The different network has a different method of community detection and we should choose a suitable method to identify communities.

**V.Conclusions**: In this paper, we review all the community mining algorithms their merits, demerits, complexity, and scale and these algorithms are divided into 3 categories. first is Traditional community mining algorithm second is community mining for overlapping community and the last is an algorithm for local community detection. The first section we discuss the introduction related to big data and community mining algorithm. The second section discusses some related work or Literature review. The third section discusses A comprehensive study of the Community detection algorithms in detail.The fourth section discusses some more community mining algorithms. The fifth section discusses General Community Detection Problems arises in big data.

## References:


[1]. Symeon Papadopoulos et al., " Community Detection in Social Media Performance and application considerations" Data Mining know Disc (2012) DOI 10.1007/s10618-011-0224-z.
[2]. Van Dongen, S., "Graph Clustering by Flow Simulation ", Ph.D. Thesis, University of Utrecht, The Netherlands. (2000).
[3]. Daniel A. Spielman et al. , " Spectral Partitioning Works Planar graphs and finite element meshes", February 13, 1996.
[4].UthayasankarSivarajahet.al."Critical analysis of big data challenges and analytical methods", "Journal of Business Research", August (2010).
[5] Ram Milan, Kamlesh Kumar Pandey et al. " Application of Graph Theory in Big Data in Digital Era", " Proceedings of National Conference on Recent Advancement in Computer Science, Mathematics, Physics & Electronics-ISBN-978-81-936440-7-2.
[6]. DongshengDuanYuhua Li et al., "Community Mining on Dynamic Weighted Directed Graphs", CNIKM 09, November 6, 2009, Hong Kong, China.
[7]. Nan Du, Bin Wu et al. " Community Detection in Large scale Social Networks", Joint 9th WEBKDD and 1st SNA-KDD Workshop '07 August 12, 2007, San Jose, California, USA.
[8]. R. Saravanakumar et al., " A Survey on the concepts and challenges of Big Data Beyond the hype", (2017).
[9]. Jaseena K.U. et al. "Issues, Challenges, And Solutions: Big Data Mining", Computer Science & Information Technology (CS &IT ).
[10]. Kamlesh Kumar Pandey, Ram Milan et al. " Mining on Relationship in Big Data era Using Apriori Algorithm", National Conference on Data Analytics, Machine Learning and Security", 15-16 February 2018, ISBN 978-93-5291-457-9.
[11]. M.EJ. Newman, M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E 69 (2) (2004) 026113.
[12].Danon L, Diaz-Guilera et al. " Comparing community structure identification", J Stat Mech-Theory E,2005
[13]. Jure Leskovec, Kevin J. Lang et al. " Empirical comparison of Algorithms for Network Community Detection Detection ", WWW.2010.
[14]. Cuijan Wang Wenzhong Tang, Bo Sun Jing Fang et. Al.," Review on Community Detection Algorithms in Social Networks", ISBN 978-1-4673-9088-0 (2015) IEEE.
[15]. Santo Fortunato. Community Detection in Graphs[j]. Physics reports 2009(3).
[16]. Danon L, Diaz-Guilera A, Duch J, et al. Comparing Communities structure identification. J Stat Mech- Theory E, 2005.
[17]. Shangfu Gong, Wanlu Chen, PengtaoJia. Survey on algorithms of community detection[j].Application Research of Computers.2013(11).
[18]. T. Hastie, R Tibshirani, J.H. Friedman, The Elements of Statistical Learning, Springer, Berlin, Germany,2001.
[19]. Rosvall, M, Bergstrom, C.T: Maps of Random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences 105(4) (2008) 1118-1123.
[20].JIERUI XIE, STEPHEN KELLEY, BOLESLAW K. SZYMANSKI Overlapping Community Detection in Networks: the state of the Art and Comparative Study. ACM Computing Surveys, Vol45, no, 2013.
[21]. I. Derenyi, G. Palla, T. Vicsek, Clique percolation in random networks, Phys. Rev. Lett. 94(16)(2005)160202.
[22]. Ahn Y Y, Bagrow J P, Lehmann S. Link communities reveal multiscale complexity in networks[j]. . Nature, 2010, 466(7307):761-764.
[23]. Gregory S. Finding overlapping communities in networks in networks by label propagation. New Journal of Physics, 2012,12 (10):102018.
[24]. Clauset A. Finding local community structure in networks [J]. Physical review E, 2005,72(2):026132.
[25].Raghavan U N, Albert R, Kumara S. Near linear time algorithm to detect community structure in large scale networks. Physical Review E, Statistical, Nonlinear, and Soft Matter Physics, 2007, 76 ( 3 pt 2): 036106
[26]. Joyce JiyoungWhang, David F. Gleich, Inderjit S. Dhillon. Overlapping community detection using seed set expansion CIKM'13.2013(2099-2108).
[27].Lancichinetti, A, Radicchi, F, Ramasco, J.J. Fortunato, S: finding statistically significantly communities in networks. Plos ONE 6(4) ( April 2011 ) e18961+
[28]. https://en.wikipedia.org/wiki/K-means_clustering
[29].Steve Gregory," Finding overlapping communities in networks by the label", Department of Computer Science, University of Bristol, Bristol BS8 1UB, England.
[30]. M.E.J Newman," Fast Algorithm for detecting Community structure in Networks", arxiv:cond-mat/0309508v1",
22 September 2003.

[31]. SerafeimTsironis et al., " Accurate Spectral Clustering For Community Detection in MapReduce Stanford.edu"
[32]. Zhao Yang et al. ," A Comparative Analysis of Community Detection Algorithms on Artificial Networks", Scientific Reports volume6, Article number: 30750 (2016
[33]. Xingyi Zhang et al., " A Fast Overlapping Community Detection Algorithm based on weak cliques for Large scale networks", "IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS, VOL. 4, NO. 4, DECEMBER 2017".
[34]. Jun Hua et al., "A New Community Detection Algorithm Based on Adding and Deleting Links", " IEEE 2 nd International Conference on Big Data Analysis" ,2017.
[35]. Mehak Mohammad," community detection in Social Networks Through Girvan Newman Algorithm"
[36]. FarnazMoradi," Improving Community Detection Methods for Network Data Analysis", Thesis for the Degree of Doctor of Philosophy,(2014).
[37]. Newman M. "Fast algorithm for detecting community structure in networks" ," Physical review E 2004,69(6):066133.doi:10.1103/PhysREvE.69.066133.
[38].Newman M. "Analysis of weighted networks" Physical review E 2004,70 (5):056131. Doi:10.1103/PhysRevE.70.056131.
[39]. Newman M. "Modularity and community structure in networks ", " Proceeding of the National Academy of Sciences 2006,103(23):8577-8582. Doi10.1073/pnas.0601602103.
[40]. Clauset A, Newman ME, " Moore C. Finding community structure in very large networks. " Physical review E 2004, 70(6):066111. Doi10.1103/PhysRevE.70.066111.
[41].Blondel VD, Guillaume JL et al. " Fast unfolding of communities in large networks", " Journal of statistical mechanics: Theory and experiment 2008. Doi:10.1088/1742-5468/2008/10/P10008.
[42].Guimera R, Sales-Pardo M, Amaral LAN, " Modularity from fluctuations in random graphs and complex networks" Physical review E 2004, 70(2):025101. Doi:10.1103/PhysRevE.70.025101.
[43]. Zhou Z et al. " Community detection based on improved modularity", " Pattern recognition 2012:638-64. Doi: 10.1007/978-3-642-33506-8_78.
[44].Duch J et al. , " Community detection in complex networks using external optimization", " Physical review E 2005, 72(2):027104. Doi:10.1103/PhyRevE.72.027104.
[45]. Ye Z et al., " Adaptive clustering algorithm for community detection in complex networks", " Physical review E 2008, 78(4):046115. Doi 10.1103/PhysRevE.78.046115.
[46]. Wahl S et al. " Hierarchical Fuzzy Spectral Clustering in Social networks Using Spectral Characterization " in the twenty eighths international flairs conference" 2015:305-310.
[47].Falkowski T et al. " A density-based community detection algorithm " " IEEE/WIC/ACM International Conference on Web Intelligence (WI); 2007:112-115. Doi 10.1109/WI.2007.74.
[48]. Nikolaev AG et al ." On efficient use of entropy centrality for social network analysis and community detection", " Social networks 2015, 40:154-162. Doi:10.1016/j.socnet.2014.10.002.
[49]. Steinhaeuser K et al. " Identifying and evaluating community structure in complex networks. "Pattern Recognition Letters 2010,31(5):413-421.doi:10.1016/J.patrec.2009.11.001.
[50]. Pizzuti C., " Ga-Net: A Genetic algorithm for community detection in social networks." " parallel problem solving form Nature-PPSN X: Springer: 2008, 1081-1090.doi:10.1007/978-3-540-87700-4_107.
[51]. PizzutiC., " A multiobjective genetic algorithm to find communities in complex networks" IEEE Transactions on evolutionary computation 2012, 16 (3):418-430. Doi:10.1109/TEVC.2011.2161090.
[52]. Hafez AI. Et al, "A genetic algorithms for community detection in social networks", " in 12 the international conference on intelligent systems design and applications (ISDA): IEEE; " 2012: 460-465. Doi:10.1109/ISDA.2012.6416582.
[53]. Mazur P et al. " A genetic algorithm approach to community detection", Acta Physica Polonica Series A- general physics 2010,117(4).
[54]. Liu X et al. " Effective algorithm for detecting community structure in complex networks based on GA and clustering " in International conference on computational science ( ICCS 07): Springer; 2007:657-664.doi:10.1007/978-3-540-72586-2_95.
[55]. Tasgin M et al. " Community detection in complex networks using genetic algorithms " arXiv preprint arXiv:0711.0491 2007.
[56]. Zadeh PM et al. " A multi population cultural algorithm for community detection in social networks", " Procedia Computer science 2015, 52:342-349. Doi:10.1016/J.procs.2015.05.105.
[57]. Nicosia V, " Extending the definition of modularity to directed graphs with overlapping communities", " Journal of Statistical Mechanics: Theory and experiment" 2009, 3, P03024. Doi:10.1088/1742-5468/2009/03/P03024.
[58]. Raghavan UN et al. " Near linear time algorithm to detect community structures in large scale network". " Physical review E 2007,76(3):036106.doi:10.1103/PhysRevE.76.036106.
[59]. Xie J et al. " Towards linear time overlapping community detection in social networks. " " Advances in Knowledge Discovery and data mining" Springer; 2012,25-36.
[60]. Hu W et al. " Finding Statistically significant communities in networks with weighted Label Propagation. ", " Social networking " 2013, 2:138-146 doi: 10.4236/sn.2013.23012.
[61]. Gregory S.," Finding overlapping communities in the network by label propagation." " NewJournal of Physics 2010"12 (10):103018.doi 10.1088/1367-2630/12/10/103018.
[62]. Xie J et al. "Labelrank: A stabilized label propagation algorithm for community detection in networks " "IEEE network science workshop (NSW):2013:138-143.
[64]. WU Z-H et al. " Balanced multi-label propagation for overlapping community detection in social networks. " Journal of computer science and technology", 2012,27(3):468-479.
[65]. Xie J et al. " Incremental community detection in dynamic networks via label propagation" " proceedings of the workshop on dynamic networks management and mining: ACM;2013:25-32.
[66]. Blei DM et al. " Latent Dirichlet allocation" " the journal of machine learning research" 2003, 3:993-1022.
[67]. Xin Y et al. " A semantic overlapping community detection algorithm based on field sampling" "expert systems with applications" 2015, 42 (1):336-375. Doi:10.1016/J.eswa.2014.07.009.
[68]. Xin Y et al. " An overlapping semantic community detection algorithm base on the ARTs multiple sampling models" " Expert systems with applications" 2015, 42(7):3420-3432.doi:10.1016/j.eswa.2014.11.029.
[69]. Xia Z et al. " Community detection based on a semantic network" " Knowledge-based systems ", 2012, 26.doi:10.1016/j.eswa.2011.06.014
[70]. Ding Y " Community detection: Topological vs. topical" " Journal of Informatics" 2011,5(4):498-514.doi:10.1016/j.joi.2011.02.006.
[71]. Ereteo G, et al. " Semtagp: semantic community detection in folksonomies", " in the proceedings of the 2011 IEEE /WIC/ACM International conference on web intelligence and intelligent agent technology-volume 01: IEEE computer society " 2011.doi: 10.1109/WIAT.2011.98.
[72]. Zhao Z et al. " Topic-oriented community detection through social objects and link analysis in social networks", " Knowledge-based systems" 2012,26:164-173. Doi:10.1016/j.knosys.2011.07.017.
[73]. Abdelbary HA et al. " A Semantic topics modeling approach for community detection social network", 2013,81(6). Doi:10.5120/14020-2177.
[74]. Deerwester SC et al. " Indexing by latent semantic analysis", Jesus 1990,41(6):391-407.
Doi:10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.
[75].Nguyen T et al. " Hyper community detection in the blogosphere" , " in proceddings of second ACM SIGMM workshop on social media" ACM;2010.
[76]. Natarajan N et al. " Community detection in content sharing social networks"," in Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining", ACM; 2013.doi:10.1145/2492517.2492546.
[77]. Xie J et al. " Overlapping community detection in networks: the state of the art and comparative study", " ACM computing surveys (CSU) 2013,45 (4):1-35.doi:10.1145/2501654.2501657.
[78]. Palla G et al. " Uncovering the overlapping community structure of complex networks in nature and society", Nature 2005,435:814-818. Doi:10.1038/nature03607.
[79]. Lancichinetti A et al. "Detecting the overlapping and hierarchical community structure in complex networks", " New Journal of physics 2009", 11 (3):033015:doi:10.1088/1367-2630/11/2/033015.
[80]. Du N et al. " Community detection in large scale social networks"," in Proceedings of the 9thWebKDD and 1st SNA-KDD 2007 workshop on web mining and social network analysis: ACM;2007:16-25.doi:10.1145/1348549.1348552.

## AUTHORS PROFILE

**Ram Milan** is pursuing a Ph.D. from Dr. HariSinghGourVishwavidyalaya (A Central University), Sagar, India, under the supervision of Prof. DiwakarShukla. Currently, He is doing research on Community Detection in Big Data using Sampling and Graph Theory. He qualified UGC NET three times. He is the author and co-author of several research papers in International journals and conference. He is a co-author of 3 books.

**Prof. DiwakarShukla** is presently working as HOD in the Department of Computer science and applications, and Dean in the School of Mathematical and Physical Sciences, Dr. HariSinghGourVishwavidyalaya, Sagar, India, and has over 25 years experience of teaching to U.G. and P.G. classes. He obtained M.Sc.(stat.), Ph.D.(stat.) degrees from Banaras Hindu University, Varanasi and served the Devi Ahilya University, Indore, M.P. as a permanent Lecturer from 1989 for nine years and obtained the degree of M.Tech.(Computer Science) from there. He joined Dr. HariSinghGourVishwavidyalaya, Sagar as a Reader in statistics in the year 1998. During Ph.D. from BHU, he was junior and senior research fellow of CSIR, New Delhi through Fellowship Examination (NET) of 1983. Till now, he has published more than 75 research papers in national and international journals and participated in more than 35 seminars/conferences at the national level. He also worked as a Professor in the Lucknow University, Lucknow, U.P., for one (from June 2007 to 2008) year and visited abroad to Sydney (Australia) and Shanghai (China) for conference participation and paper presentation. He has supervised fourteen Ph.D. theses in Statistics and Computer Science and seven students are presently enrolled for their doctoral degree under his supervision. He is the author of two books. He is a member of 11 learned bodies of Statistics and Computer Science at the national level. The area of research he works for are Sampling Theory, Graph Theory, Stochastic Modeling, Data mining, Big Data, Operation Research, Computer Network, and Operating Systems.

**Mr. Kamlesh Kumar Pandey** is pursuing a Ph.D. from Dr. HariSinghGourVishwavidyalaya (A Central University), Sagar, India, under the supervision of Prof. DiwakarShukla. Currently, He is doing research on the design of Big Data Mining algorithms with respect to three dimensions of Big Data. He is the author and co-author of several research papers in International journals and conference such as IEEE, Springer, and others. He has 6 years of teaching and research experience. He awarded Training of Young Scientist in 34th M.P. Young Scientist Congress.