6. OVERLAPPING COMMUNITY DETECTION USING CLIQUE PERCOLATION

Community structures are defined as the division of network into various modules like groups, clusters, communities which are connected to each other. The modules comprises of nodes and edges which have dense connections between the nodes within the same modules but have sparse connections between nodes in the different modules. Graph theory techniques are adopted to create these modules. Graph theory emphasizes on finding communities in a network using different algorithms and optimizes the solution. There are number of approaches and tools available to generate community structure. The subgraph analysis with maximal k-core, k-plex and maximal k-clique described in the previous chapter has the ability to find largest subgroups but inefficient to detect weak sub communities and overlapping communities of the given network. These overlapping communities are needed to identify inter and intra group interaction between the various nodes in the network. This chapter demonstrates overlapping community detection and elucidates the implementation of clique percolation method. The importance of groud truth communities is also presented in this chapter.

6.1 INTRODUCTION

Community detection is an important tool to analyze hidden information such as functional module and topology structure in complex networks. Compared with traditional community detection, it is more demanding to find overlapping communities in complex networks, especially when the networks are of large scales. Among various overlapping community detection techniques, the well-known Clique Percolation Method (CPM) has shown promising performance in terms of quality community detection, but suffers from serious curse of dimensionality due to its high computational complexity, which makes it very unlikely to be applied to large-scale networks.

The node based overlapping community detection algorithms divide nodes of the network into different communities directly, utilizing the structure information of nodes. However, traditional node clustering and relatively proposed clustering methods have inherent drawbacks to discover overlapping communities. Node clustering is inadequate to capture the pervasive overlaps, while link clustering is often criticized due to the high computational cost and ambiguous definition of communities. Overlapping community detection is still a dreadful challenge.

The clique percolation method is used to investigate the changing spatial organization of the network. Non overlapping community detection algorithms assign nodes into exclusive communities and, when results are mapped, these techniques may generate spatially disjointed geographical regions, an undesirable characteristic for geographical study. Alternative overlapping community detection algorithms permit overlapping membership where a node can be a member of different communities. Such a structure simultaneously accommodates spatial proximity and spatial separation which happen with respect to a node in relation to other nodes in the system. Applying such a structure in geographical analysis assists preserve well-established principles regarding spatial relationships within the geography discipline. The result can also be mapped for display and correct interpretation.

The CPM is chosen in this study due to the complete connection within cliques which enables the formation of highly connected networks. However, the CPM has been shown to be among the best performing overlapping community detection algorithms. Detecting communities in a network only exposes certain characteristics of the spatial organization of the network, rather than providing explanation of the spatial-network patterns revealed. Full interpretation of the pattern should rely on the attribute data and additional information. This illustrates the value of an integrated approach in geographical analysis using both social network analysis and spatial analysis techniques [87].

6.2 OVERLAPPING COMMUNITY DETECTION ALGORITHM

This section describes clique finding algorithms which are used for detecting overlapping communities in networks. In many real-world networks, it is natural to find vertices or members that belong to more than one group or community at the same time. Most community detection algorithms designed to identify mutually independent communities in a large network required the division of networks into smaller connected clusters by the removal of key edges which connect dense sub-graphs, and therefore, are not suitable for detecting overlapping communities. Clique percolation is an effective algorithm for detecting overlapping communities in large graphs.

Clique Percolation Method

The Clique Percolation Method (CPM) is one effective approach for detecting overlapping communities in a network. The basic principle of CPM is that a typical community is likely to be made up of several cliques that share many of their vertices. A clique of size k is called a k-clique, and two k-cliques are called adjacent if they share k−1nodes.A k-clique community is a union of all k-cliques that can be achieved from each other through a series of adjacent k-cliques. A method is devised to extract such k-clique communities of a network. K-clique communities allow overlaps, i.e., common vertices shared by the communities. Fig.6.1 illustrates an example of overlapping community detection by CPM.

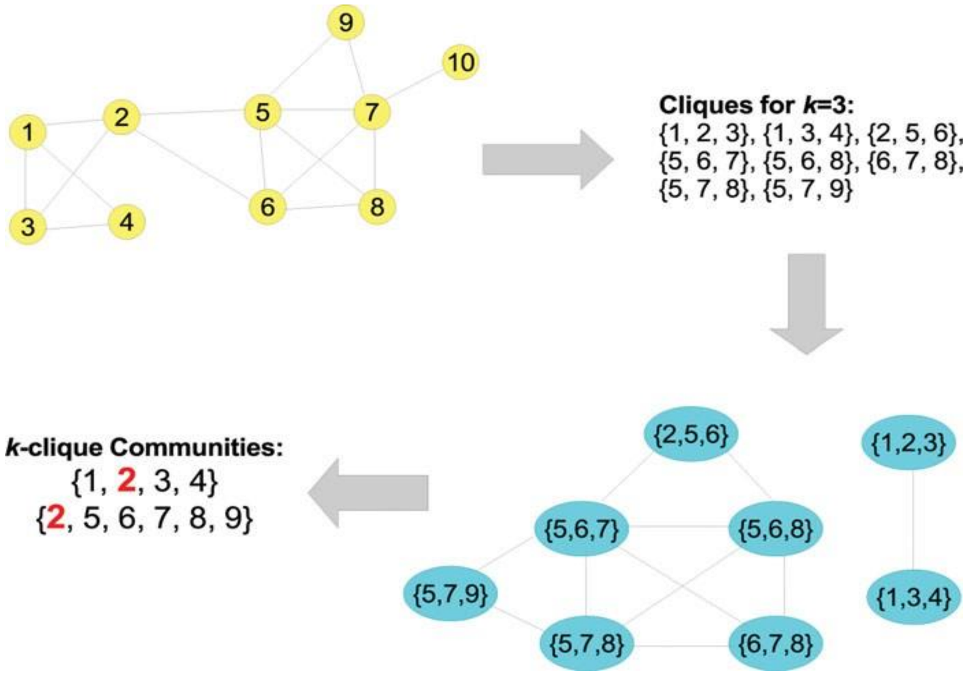

Fig. 6.1 Illustration of Overlapping Community Detection by CPM

Let n vertices of the input graph G = (V, E) are denoted as {v_1, v_2,.,v_n}. The set of vertices adjacent to a vertex v_i, i.e. the set of its neighbors, is denoted by N (v_i) and the degree of the vertex v_i, is denoted by d (v_i). In this algorithm, the degree is computed once for each vertex at the beginning. Given a graph, first it extracts all cliques of size k. This is followed by generating the clique graph, in which each k-clique in the original graph is represented by a vertex. An edge is then added between any two k-cliques in the clique graphs that are adjacent. For example, if k = 3, an edge is added between any two 3-cliques in the clique graph that share two common vertices. The connected components in the clique graph represent a community, and the actual members of the community are obtained by gathering the vertices of the individual cliques that form the connected component. In this manner, it obtains two communities, which share a common vertex, forming an overlapping community structure.

Also, the algorithm that tries to locate the k-cliques individually and determines the adjacency between them can be very slow for large networks. Hence largest k-cliques are determined in this clique percolation algorithm. A large clique of size q ≥ k contains $\binom{q}{k}$ different k-cliques. Two observations are made that helps one comes up with a better strategy. Two rules are followed. First, a clique of size q is clearly a k-clique connected subset for any k ≤ q. Second, two large cliques that share at least k − 1 nodes form one k-clique connected component. Leveraging these two rules, the strategy of searching for k-cliques individually is avoided and instead first locates the largest cliques in the network and then looks for the k-clique connected subsets of given k by learning the overlap between them. The algorithm first constructs a symmetric clique-clique overlap matrix, in which each row represents a large clique, each matrix entry is equal to the number of common nodes between the two corresponding cliques, and each diagonal entry is equal to the size of the clique.

Cliques are fully connected sub-graphs of k vertices. K-clique adjacency means two k-cliques are adjacent if they share k-1 vertices. A k-clique chain is a sub-graph which is the union of a sequence of adjacent k-cliques. Two k-cliques are k-clique connected if they are elements of a k-clique chain. A k-clique percolation cluster or component is a maximal k-clique connected sub-graph, i.e. it is the blending of all k-cliques that are k-clique-connected to a particular k-clique. The k-clique communities for a given k are then equivalent to connected clique components in which the neighbouring cliques are linked to each other by at least k − 1 common nodes. These components are then found by erasing every off-diagonal entry smaller than k−1 and every diagonal entry smaller than k in the matrix, replacing the remaining elements by one, and then carrying out a component analysis of this matrix [88].

The algorithm finds k-cliques, which correspond to fully connected sub-graphs of k nodes. It determines a community as the maximal union of k-cliques that can be reached from each other through a chain of adjacent k-cliques. First, all of the existent maximal k-clique percolation clusters for the given k are discovered. The k-clique percolation cluster is a maximal k-clique associated sub-graph. This is the union of all k-cliques that are k-clique-connected to a particular k-clique. The percolation transition takes place when the probability of two vertices being connected by an edge reaches the threshold $p_c(k) = [(k − 1)N] −1/(k−1)$.

This is because only small clusters are expected for any k at which the networks is below the transition point, but large clusters also appear, which corresponds to locally dense structures.

The process flow of Clique Percolation Algorithm (CPA) for finding overlapping community is given below.

Input: The network, G and the clique size, k

Output: Community structure, C

- The network, G and the clique size, k
- K-clique is a clique with k nodes where a clique is a complete sub graph.
- Several K-cliques communities are formed from complete network
- From the network K-cliques community forms a union of all k-clique
- A union of k-clique is formed which can be attained from each other through a series of adjacent k-cliques.
- If and only if two k-cliques are sharing k-1 nodes only than it is said to be adjacent k-cliques.

The most computationally expensive section in this algorithm is the clique-graph creation process. It achieves an extensive search on the space of cliques, looking for couples that share k-1 nodes. In the basic implementation there are two nested for loops comparing cliques and then performing n*n iterations [89]. There are two variants of clique percolation method namely Optimized Clique Percolation Method (OCPM) and Parallel Clique Percolation Method (PCPM) which are mentioned below.

Optimized Clique Percolation Method

OCPM is an optimized approach which discovers couples of cliques in the search space. The algorithm executes an exhaustive search of couples that share k-1 nodes. The two nested for loops over the same list of cliques correspond to a symmetric matrix-based search space can be reduced in investigating either the upper or lower part [90]. This implementation therefore reduces the number of iterations to n*(n-1)/2.

Parallel Clique Percolation Method

PCPM parallelizes the search of couples of cliques exploiting the number of cores that CPU have at their disposal. It also requires defining the number of clusters to parallel the execution. The algorithm is parallelized and show performance results on a shared-memory platform. The i^{th} iteration of the for loop in the algorithm computes the size of the largest clique that contains the vertex v_i. During such a concurrent computation, different processes discover maximum cliques of different sizes. For the pruning steps to be most effective, the current globally largest maximum clique size is communicated to all processes as soon as it is discovered. In a shared-memory programming model, the global maximum clique size is stored as a shared variable accessible to all the processing units, and its value is updated by the relevant processor at any given time [91, 92].

6.3 OVERLAPPING COMMUNITY DETECTION MODEL

The overlapping community detection model has been constructed with three components (i) input (ii) process and (iii) output. The input component uses twitter network data presented in chapter 3. The graph representation of the twitter data is used as input. The second component incorporates overlapping community detection process wherein three different algorithms described in sections 6.2 are employed. The process logic uses Clique Percolation Method (CPM) and its variants such as Optimized Clique Percolation Method (OCPM), Parallel Clique Percolation Method (PCPM) to find overlapping communities. In the first case clique percolation clique algorithm is implemented using node list representation of the given network. Overlapping community detection process is performed by determining the complete graphs and combining those complete graphs sharing k-1 nodes to form communities. In the second case, optimized clique percolation method is employed wherein the communities are searched by optimizing the search space with respect to a global quantity from the given graph. Parallel clique percolation algorithm is implemented in the third case wherein the overlapping community is identified by parallelizing the execution of clusters using node list. The third component is the output logic in which the effectiveness overlapping community detection algorithms is evaluated using ground truth data with respect to evaluation measures and also it analyses the detected communities based on their membership distribution. The inferences are drawn based on the experiment results. The architecture of the model is shown in Fig. 6.2.

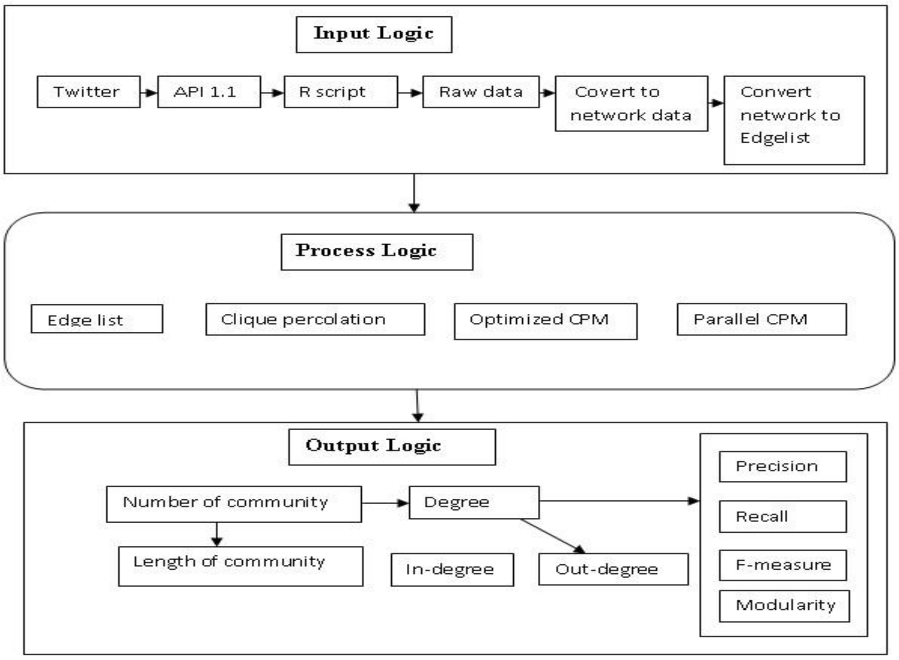

Fig. 6.2 Overlapping Community Detection Model

6.4 GROUND TRUTH COMMUNITIES AND EVALUATION

Community detection has become one of the most challenging and studied problems in complex network analysis, due to its relevance for a wide range of applications such as the study of information and disease spreading, the prediction of future interactions and activities of individuals, and even the analysis of the patterns of human mobility. Nodes in real-world networks arrange into densely linked communities where edges show with high concentration among the members of the community. Identifying such communities of nodes has proven to be a challenging task mainly due to abundance of definitions of a community, intractability of algorithms, issues with evaluation and the lack of a reliable gold-standard ground-truth.

In literature, the most commonly used evaluation method is to compare the community set produced by an algorithm on a network with ground truth communities of the same network. The evaluation procedure is clear-cut only if the ground-truth set of communities is known. Due to the scares availability of real networks with ground truth communities, the evaluation of an algorithm is often performed using synthetic network generators that also provide ground truth communities.

A novel community evaluation approach that leverages ground truth communities has been adopted to provide valuable insights on the quality of the community sets produced by a community detection algorithm. The four classes of structural definitions used in determining ground truth communities are (1) internal community connectivity, (2) external connectivity of the nodes (3) both internal and external community connectivity (4) network modularity. The standard evaluation metrics such as precision, recall, F-measure are used to determine the performance of community detection approaches. These measures are computed by comparing the predicted communities against the ground truth communities of the network. Ground-truth communities are a real group that defines connectivity patterns of network communities. A good structural definition of a community would be such that it would detect connectivity patterns that correspond to the ground-truth. Based on set cardinality, evaluation measures such as precision and recall are defined to find the overlaps nodes between pairs of communities. The proportion of correctly assigned nodes is known as purity. Each identified community is matched to the one with the maximum overlap in the reference one, and then the accuracy of this assignment is measured by counting the number of correctly assigned nodes.

Two measures, namely community precision and community recall, which provide information about how much the nodes of a given community tend to be in the same ground truth community are defined. In particular, community precision quantifies the level of label homophiles between a community and a ground truth community, while the community recall quantifies the ratio of nodes in the ground truth community enclosed by a given algorithm community. To validate the first, it computes the proposed community precision and community recall metrics on the produced community sets in order to compare them on the ground truth. Given an algorithmic community, precision indicates how many vertices are actually in the same ground truth community. Given a ground-truth community, recall indicates how many vertices are predicted to be in the same community in a retrieved community.

Given a community set X produced by an algorithm and the ground truth community set Y, for each community x ∈ X, the nodes are labelled with the ground truth community y ∈Y they belong to. Then it matches community x with the ground truth community with the highest number of labels in the algorithm community. This procedure produces (x, y) pairs having the highest similarity between the node labels in x and all the ground truth communities. The quality of the mappings is measured by the two following measures [90].

Precision - the percentage of nodes in x labelled as y, computed as

$$P = \frac{|x \cap y|}{|x|} \in [0, 1] \qquad (6.1)$$

Recall - the percentage of nodes in y covered by x, computed as

$$R \ = \frac{|x∩y|}{|y|} ∈ [0, 1] \qquad (6.2)$$

Given a pair (x, y) the above two measures describe the overlap of the members in communities. A perfect match is acquired when both precision and recall are 1. It has a many-to-one mapping i.e., multiple communities in X can be connected to a single ground truth community in Y. This methodology of evaluating the algorithms producing overlapping communities is more appropriate. The evaluation can be also summarized into a single number, the harmonic mean of precision and recall, obtaining the F1-measure which provides a clear and concise evaluation of the quality score of a community set.

$$F1 = 2\frac{precision * recall}{precision + recall} \qquad (6.3)$$

The mean F1, along with its standard deviation, makes possible to compare the performances of different algorithms on the same network with ground truth communities. Ground truth communities obtained manually using the results of sub graph analysis for the sample twitter network is given below.

(1,2,1,3,1,4,1,7,1,10,1,12,1,13,1,14,1,16,1,17,1,18,1,23,1,24,1,25,1,28,1,29,1,30,1,33,1,34,1,35,1,38,1
,40,1,41,1,43,1,44,1,45,1,46,1,47,1,48,1,49,1,50,1,51,1,53,1,55,1,56,1,57,1,59,1,62,1,63,1,64,1,65,1,6
6,1,67,1,68,1,71,1,72,1,73,1,74,1,75,1,76,1,80,1,81,1,82,1,83,1,84,1,85,1,86,1,89,1,90,1,91,1,92,1,93,
1,94,1,95,1,97,1,98,1,99,1,100,1,103,1,104,1,105,1,106,1,107,1,108,1,109,1,112,1,113,1,114,1,116,1,
119,1,120,1,121)

(2,1,2,3,2,4,2,5,2,7,2,8,2,9,2,11,2,12,2,13,2,14,2,15,2,16,2,17,2,18,2,19,2,20,2,22,2,23,2,26,2,27,2,30,
2,31,2,34,2,35,2,38,2,39,2,42,2,43,2,46,2,47,2,50,2,51,2,52,2,54,2,55,2,58,2,62,2,63,2,66,2,67,2,70,2,
71,2,74,2,75,2,77,2,79,2,82,2,83,2,86,2,87,2,90,2,91,2,94,2,95,2,98,2,99,2,102,2,103,2,106,2,107,2,11
0,2,111,2,114,2,117,2,118,121)

(3,1,3,5,3,9,3,13,3,17,3,18,3,19,3,20,3,21,3,23,3,26,3,27,3,30,3,31,3,34,3,35,3,37,3,38,3,39,3,42,3,43,
3,46,3,47,3,48,3,50,3,51,3,52,3,54,3,55,3,58,3,59,3,62,3,63,3,66,3,67,3,70,3,71,3,74,3,75,3,79,3,82,3,
83,3,86,3,87,3,90,3,91,3,94,3,95,3,99,3,102,3,103,3,106,3,107,3,110,3,111,3,114,3,117,3,121)

(50 96 114 74 120 25 181 89 14 36 16 9 10 76 163 76 148 86 13 71 173 82 14 66 163 71 152 68 17
55 165 84 138 40 112 59 152 98 115 30 113 58 116 91 129 96 126 98 142 88 115 61 188 98 11 28
146 79 131 65 138 50 183 85 11 54 158 65 13 52 126 54 170 94 112 69 159 71 16 23 137 91 115 83
17 22 132)

(44 129 37 127 93 113 70 188 93 152 67 17 44 131 51 128 87 179 85 14 73 13 27 140 73 14 87 110
14 12 99 118 97 164 95 113 51 111 15 176 87 15 93 145 54 11 21 17 91 183 90 147 64 118 81 123 84
145 71 126 49 132 68 119 28 170 81 171 93 11 33 19 65 136 74 128 63 118 30 10 6 129 62 123 40
136 80)

(117 24 141 43 163 88 111 46 127 72 172 80 111 66 149 82 165 88 137 71 120 89 19 20 114 19 110
19 116 66 158 60 156 89 119 79 120 71 174 87 157 99 18 69 16 72 152 81 10 59 120 54 160 93 123
75 147 62 121 55 14 17 118 84 127 43 133 60 143 63 137 87 12 91 126 60 111 39 114 18 114 92 154
91 11 77)

(115 63 141 54 120 96 19 29 16 87 115 19 139 54 165 75 147 86 158 90 113 25 11 10 132 82 192 99
146 70 111 96 168 87 16 98 166 99 163 96 162 73 161 80 112 72 124 94 181 83 139 64 110 97 18 33
163 78 148 72 111 92 133 47 140 62 130 49 134 44 163 84 118 43 189 92 167 76 113 56 144 58 155
72 147 95 144 86 117 88 178 96 150 63 122 31 159 99 157 87 14 7 140 64 12 51 10 47 114 83 153 66
13 42 11 46)

(137 89 136 72 15 32 136 69 16 47 173 89 154 85 11 95 134 39 149 69 178 95 111 43 141 95 131 32
16 65 136 47 128 79 19 17 181 95 155 65 129 98 145 52 151 97 166 91 131 97 119 72 12 16 146 80
147 66 168 69 11 13 167 97 180 87 143 93 132 70 121 25 182 99 110 90 125 34 16 12 120 32 137
66)

(60 15 41 112 84 19 46 124 82 125 68 16 38 139 76 136 82 110 45 117 30 132 98 111 32 19 48 174
78 133 84 13 70 136 56 151 64 143 47 18 27 120 91 16 62 13 65 129 56 116 68 18 97 126 28 10 7 1
14 59 118 70 11 6 157 97 121 75 119 55 175 95 133 65 146 61 196 97 159 87 110 69 157 75 158 91
121 53)

(145 88 134 95 135 87 183 93 181 88 121 31 151 94 12 89 12 8 163 67 148 83 125 69 118 60 172 89
17 99 11 67 17 58 165 87 111 63 172 82 174 83 113 76 13 85 139 48 113 61 111 21 144 61 10 30 116
81 127 48 110 22 134 41 146 68 118 91 11 49 157 76 121 62 14 26 151 95 163 83 160 85 15 26 119)

6.5 EXPERIMENTS AND RESULTS

In this work, the next level of investigation of communities has been performed using three overlapping community detection algorithms such as clique percolation, optimized clique percolation and parallel clique percolation algorithm have been implemented. The same twitter network is used, and the experiments have been carried out in R 3.5.1.

Results of Clique Percolation

Clique percolation algorithm has been implemented using edge list of the sample network. CPM overlapping algorithm discovered 198 communities in the network out of which eighty-nine communities around 180 members in the community network. 89 communities are having the large number of nodes accounting to 1800 to 501 sizes of the nodes in the community. Seventy-seven communities are having the medium number of nodes that is 500 to 101 in the community. Thirty-two communities are having the small number of nodes and community sizes ranging from 20 to 100 in the network. The modularity score obtained is 0.77. The sizes of the overlapping communities are established for each community in the network. The dense community size denotes that the friends and followers are more interactive with community of the network and shares more information between each node. When the size of the community is sparse, the friends and followers are less interactive within community. Out of 198, there are 166 dense communities and 32 sparse communities detected. The size of the largest community obtained is 1690 and the size of the smallest subgroup is 49. The communities detected by CPM method from the sports person’s network are shown in Fig. 6.3.

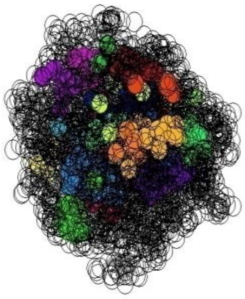

Fig. 6.3 Communities detected of CPM

A sample of 10 communities with node ids is given below. The membership distribution of ten communities is shown in Table XXIII and illustrated in Fig. 6.4.

[1] 91 100 2295 2294 2287 2286 335 66 92 1817 1814 1803 73 104 1960 1959 1956 1955 1954 1953 1951 1950 1949 1947 1946 1945 1944 1658 1356 1308 887 796 717 66 92 1817 1814 1803 1191 1190 1188 1187 1186 1185 1184 1180 1176 1111

[2] 74 103 1969 1966 1965 1377 726 725 724 722 189 66 92 1817 1814 1803 1191 1190 1188 1187 1186 1185 1184 1180 176 1111

[3] 73 104 1960 1959 1956 1955 1954 1953 1951 1950 1949 1947 1946 1945 1944 1658 1356 1308 887 796 717 1698 1191 1190 1188 1187 1186 1185 1184 1180 1176 1111

[4] 67 101 1834 66 92 1817 1814 1803 1698 1191 1190 1188 1187 1186 1185 1184 1180 1176 1111

[5] 66 92 1817 1814 1803 1191 1190 1188 1187 1186 1185 1184 1180 1176 1111

[6] 61 83 1698 1191 1190 1188 1187 1186 1185 1184 1180 1176 1111

[7] 49 79 92 1817 1814 1803 66 92 1817 1814 1803 1191 1190 1188 1187 1186 1185 1184 1180 1176 1111 1413

[8] 39 50 1202 1201 1198 1194 1191 190 1188 1187 1186 1185 1184 1180 1176 1111 1698 1191 1190 1188 1187 1186 1185 1184 1180 1176 1111

[9] 35 38 1098 1191 1190 1188 1187 1186 1185 1184 1180 1176 1111

[10] 30 93 970 966 963 961 960 959 957 956 955 953 951 950 947 944 943

Table XXIII Sizes of CPM Overlapping Communities

Overlapping Communities	Size of Communities
CPM 1	567
CPM 2	1280
CPM 3	1690
CPM 4	699
CPM 5	904
CPM 6	1009
CPM 7	1460
CPM 8	1739
CPM 9	833
CPM 10	1203

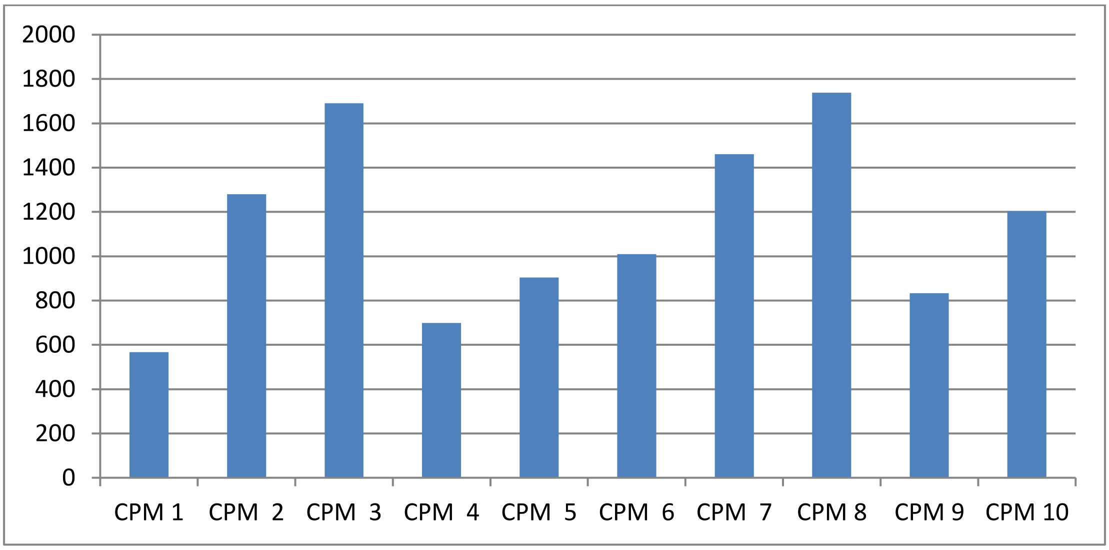

Fig. 6.4 Size of CPM Overlapping Community

Also, the degree measures are evaluated using clique percolation algorithm and the results for a sample of 10 communities are presented in Table XXIV. From the results, it is observed that in-degree of 90 communities lies between 501 to 1800 and the in-degree of 76 communities lies between 101 to 500 which indicate that friends and followers are more interactive with other nodes. The in-degree of 32 communities lies between 20 to 100, which show less interaction with other nodes because it is very popular node in the network. The high out-degree of 96 communities lies between 101 to 250. High out-degree value of 96 communities suggests more interaction from outer node to these nodes. For the remaining 102 communities, the out-degree lies in the range of 20 to 60. The in-degree and out-degree of all 198 communities of the sample input network is illustrated in Fig. 6.5.

Table XXIV Degree Measures of CPM Communities

CPM Communities	1	2	3	4	5	6	7	8	9	10
In-degree	1100	861	698	439	794	813	362	783	1661	830
Out-degree	65	48	69	110	47	68	89	127	147	99

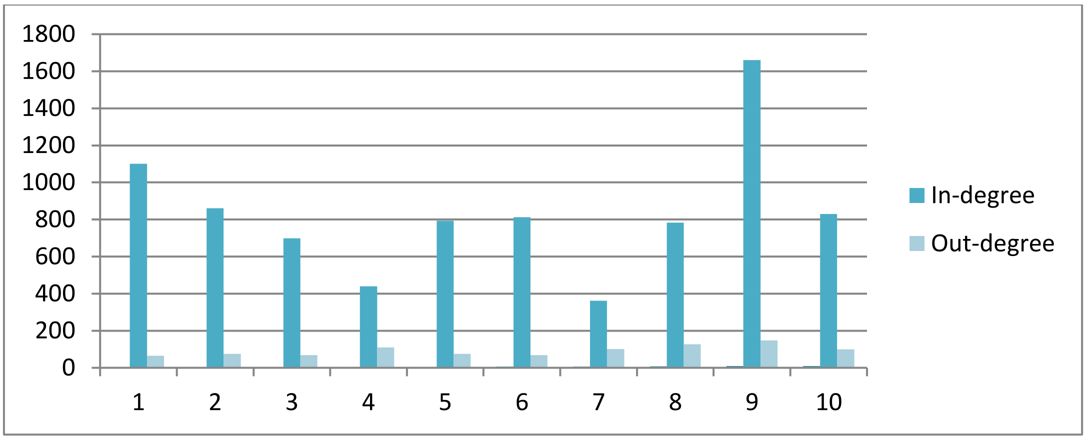

Fig. 6.5 In-Degree and Out-Degree of Overlapping Communities by CPM

As there is no direct method for evaluating the performance of overlapping community detection algorithm., ground truth communities are compared against the predicted communities and various measures such as precision, recall, F-measure are evaluated as described in section 6.4. The performance of the overlapping community detection CPM evaluated with respect to metrics such as precision, recall, F-measure yielded the results of precision as 0.76 whereas the recall and the F-measure is 0.57and 0.73 respectively. The results of various analytical measures are tabulated in Table XXV.

Table XXV Analytical Measures of CPM Communities

Number of Communities	198
Dense Communities	166
Sparse Communities	32
Size of Largest Community	1690
Size of Smallest Community	49
Largest In-Degree	1698
Largest Out-Degree	204
Modularity Score	0.7725246
F measure	0.73
Precision	0.76
Recall	0.57

Results of Optimized Clique Percolation

The OCPM algorithm reduces the number of iterations for detecting overlapping communities and also decreases the computational time. OCPM overlapping algorithm discovered 180 communities in the network. Seventy-six communities are having the large number of nodes accounting to 1800 to 501 sizes of the nodes in the community of the network. These communities are dense and share more information between each node. Sixty-three communities are having the medium number of nodes ranging from 500 to 101 in the community of the network. Incoming and out coming nodes are interactive between each community. Forty-one communities are having the small number of nodes and community sizes are 20 to 100 in the network. So, this network has large numbers of nodes sharing the link for each node. It is concluded that this community has less interaction between nodes on the network. Fig.6.6 shows overlapping communities detected by OCPM from the sports person's twitter network.

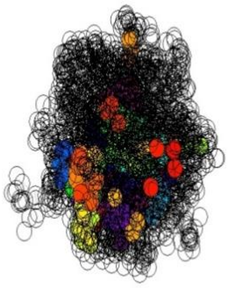

Fig. 6.6 OCPM Community Size and Network

A sample of 10 communities with node ids is given below. The membership distribution of ten communities is shown in Table XXVI and illustrated in Fig. 6.7

[1] 24 36 798 794 793 792 779 778 777 776 775 774 9 135 134 131 130 129 127 126 120 119 118 113 112 111 107 106

[2] 22 31 737 11 55 435 433 430 426 425 422 419 418 417 416 415 414 412 411 410 409 408 407 406 405 377 20 42 695 694 692 690 689 687 686 685 683 682 680 679 677 676 674 673 672 671 371 137

[3] 21 78 707 20 42 695 694 692 690 689 687 686 685 683 682 680 679 677 676 674 673 672 671
137 9 135 134 131 130 129 127 126 120 119 118 113 112 111 107 106 371

[4] 20 42 695 694 692 690 689 687 686 685 683 682 680 679 677 676 674 673 672 671 371 137

[5] 19 100 667 666 664 660 659 658 655 654 652 646

[6] 17 18 607 606 605 604 597 321 210 411 410 409 408 407 406 40

[7] 14 96 526 525 524 518 514 501 499 411 410 409 408 407 406 40

[8] 13 27 493 492 490 486 483 250 683 682 680 679 677 676 674

[9] 11 55 435 433 430 426 425 422 419 418 417 416 415 414 412 411 410 409 408 407 406 405

[10] 6 68 295 293 286 279 273 272 683 682 680 679 677 676 674

[11] 5 34 262 259 256 252 244 242 239 234 144 683 682 680 679 677 676 674

[22] 3 37 196 192 190 188 183 179 176 175 172 683 682 680 679 677 676 674

[23] 42 88 168 166 164 160 159 157 155 154 148 145 143 683 682 680 679 677 676 674

[24] 1 9 135 134 131 130 129 127 126 120 119 118 113 112 111 107 106 377

The modularity score obtained is 0.79. The different sizes of the overlapping communities are established for each community in the network. The interaction among friends and followers in community is high for a dense community. When the size of the community is sparse, the friends and followers are less interactive within community. Out of 180, there are 129 dense communities and 51 sparse communities detected. The size of the largest community obtained is 1710 and the size of the smallest subgroup is 51.

Table XXVI Sizes of OCPM Overlapping Communities

Overlapping Communities	Size of Communities
OCPM 1	1360
OCPM 2	936
OCPM 3	1720
OCPM 4	1368
OCPM 5	1100
OCPM 6	628
OCPM 7	1674
OCPM 8	399
OCPM 9	732
OCPM 10	549

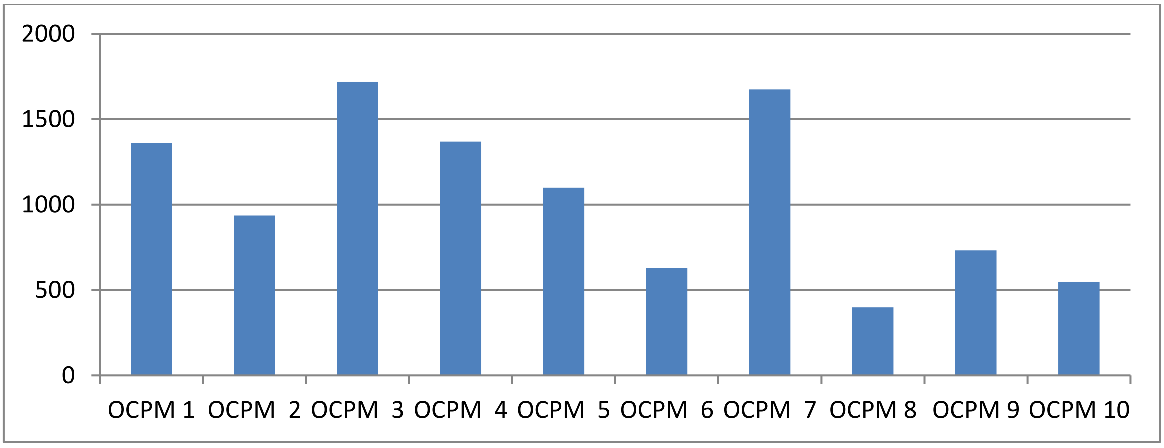

Fig. 6.7 Sizes of OCPM Overlapping Communities

Also, in-degree of 97 communities lies between 501 to 1800 and the in-degree of 32 communities lies between 101 to 500 which indicate that friends and followers are more interactive with other nodes. The in-degree of 51 communities lies between 20 to 100, which show less interaction with other nodes because it is very popular node in the network. The high out-degree of 98communities lies between 101 to 250. High out-degree value of 98 communities suggests more interaction from outer node to these nodes. For other 82 communities, the out-degree lies in the range of 20 to 60. The degree measures of overlapping community detection algorithm are evaluated using optimized clique percolation algorithm and the results for a sample of 10 communities are presented in Table XXVII. The in-degree and out-degree of all 180 communities of the sample input network is illustrated in Fig. 6.8.

Table XXVII Degree Measures of OCPM Communities

CPM Communities	1	2	3	4	5	6	7	8	9	10
In-degree	1101	861	1213	1431	1791	890	603	708	1764	583
Out-degree	155	145	69	140	87	98	99	120	167	194

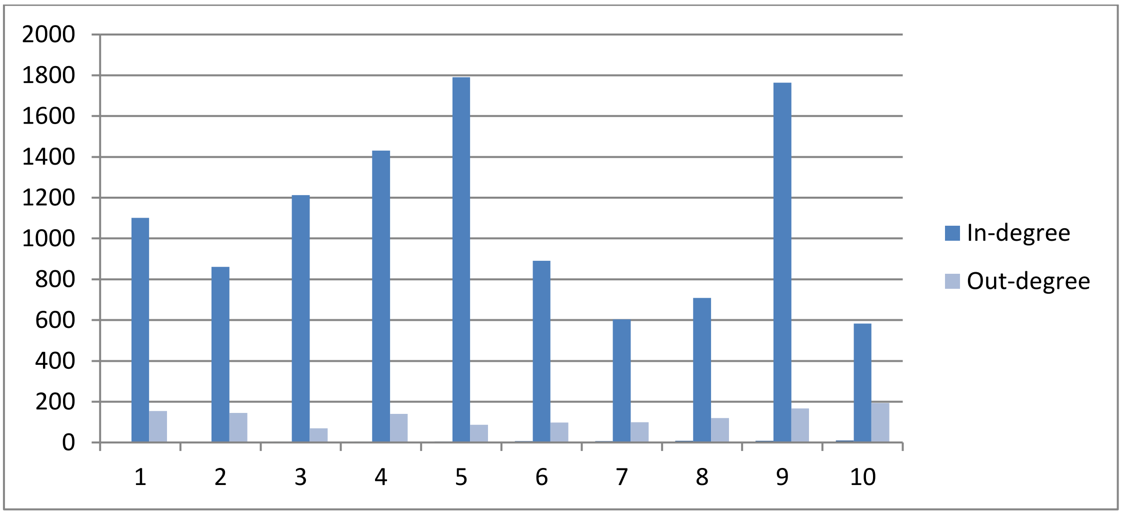

Fig. 6.8 In-Degree and Out-Degree of Overlapping Community OCPM

The performance of the optimized overlapping community detection algorithm is evaluated for its precision, recall, F-measure of predicted communities against ground truth data. The precision obtained is 0.78 whereas the recall and the F-measure is 0.58 and 0.75 respectively. The results of various analytical measures are tabulated in Table XXVIII.

Table XXVIII Analytical Measures of OCPM Communities

Number of Communities	180
Dense Communities	129
Sparse Communities	51
Size of Largest Community	1710
Size of Smallest Community	51
Largest In-Degree	1791
Largest Out-Degree	214
Modularity Score	0.79
F measure	0.75
Precision	0.78
Recall	0.58

Results of Parallel Clique Percolation

In PCPM simultaneous computation of discovering maximum cliques of different sizes is done and hence reduces the computational time. PCPM discovered 170 communities in the network of which 74 communities have large number of nodes that is 1800 to 501 nodes in the community of the network which indicate that the communities are dense and share more information between each node. Fifty-three communities are having the medium number of nodes accounting to 500 to 101 in the community of the network. Forty-three communities are having the small number of nodes and community sizes ranging from 20 to 100 in the network. So, this network has large numbers of nodes sharing the link for each node. Fig.6.9 shows the cricket player's PCPM overlapping community network.

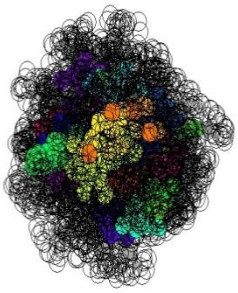

Fig. 6.9 PCPM Overlapping Community Network

A sample of 10 overlapping communities detected by parallel optimized clique percolation is given below.

```
[1] 5 34 262 259 256 252 244 242 239 234 144 683 682 680 679 677 676 674
[2] 3 37 196 192 190 188 183 179 176 175 172 683 682 680 679 677 676 674
[3] 2 88 168 166 164 160 159 157 155 154 148 145 143 683 682 680 679 677 676 674
[4] 1 9 135 134 131 130 129 127 126 120 119 118 113 112 111 107 106 377
[5] 37 196 192 190 188 183 179 176 252 244 242 239 234 144 683
[6] 14 160 159 157 155 154 148 145 96 526 525 524 518 514 501 499 411 410 409 408 407 406 40
[7] 13 27 160 159 157 155 154 148 145 493 492 490 486 483 250 683 682 680 679 677 676 674
[8] 21 78 707 20 42 695 694 692 690 689 687 686 685 683 682 680 679 677 676 674 673 672 671
[9] 21 78 707 20 42 695 694 692 690 689 687 686 685 683 682 680 679 677 676 674 673 672 671
[10] 135 134 137 139131 130 129 127 126 120 119 118 113 112 111 107 106 371 689 687 686 685
```

The modularity score obtained is 0.84. The different sizes of the overlapping communities are established for each community in the network. The overlapping community detection method discovered different size of dense and sparse communities in the network. Out of 170, there are 148 dense communities and 22 sparse communities detected. The size of the largest community obtained is 1790 and the size of the smallest subgroup is 53. The results for 10 PCPM overlapping communities in the given network are presented in Table XXIX and illustrated in Fig. 6.10.

Table XXIX Sizes of PCPM Overlapping Communities

Overlapping Communities	Size of Communities
PCPM 1	1450
PCPM 2	1324
PCPM 3	1756
PCPM 4	1347
PCPM 5	1249
PCPM 6	1003
PCPM 7	1659
PCPM 8	1124
PCPM 9	986
PCPM 10	994

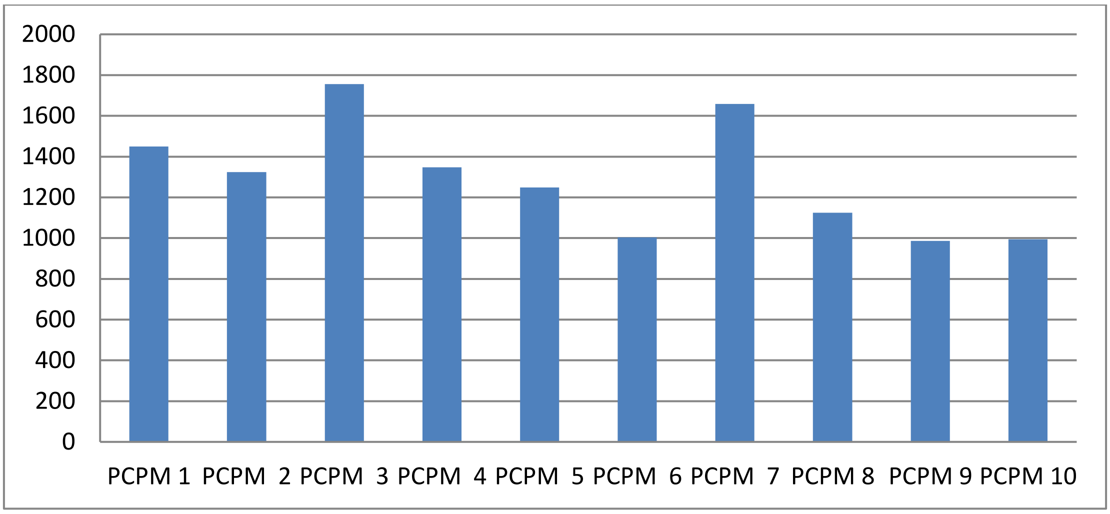

Fig. 6.10 Sizes of PCPM Overlapping Community

Also, in-degree of 107 communities lies between 501 to 1800 and the in-degree of 41 communities lies between 101 to 500 which indicate that friends and followers are more interactive with other nodes. The in-degree of 22 communities lies between 20 to 100, which show less interaction with other nodes because it is very popular node in the network. The high out-degree of 123 communities lies between 101 to 250. High out- degree value of 123 communities suggests more interaction from outer node to these nodes. For other 47 communities, the out-degree lies in the range of 20 to 60. The degree measures of overlapping community detection algorithm are evaluated using parallel clique percolation algorithm and the results for a sample of 10 communities are presented in Table XXX and are illustrated in Fig.6.11.

Table XXX Degree Measures of PCPM Communities

PCPM Communities	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
In-degree	1315	1861	1313	1000	1079	804	703	718	1746	1583
Out-degree	165	135	169	170	96	198	94	220	176	144

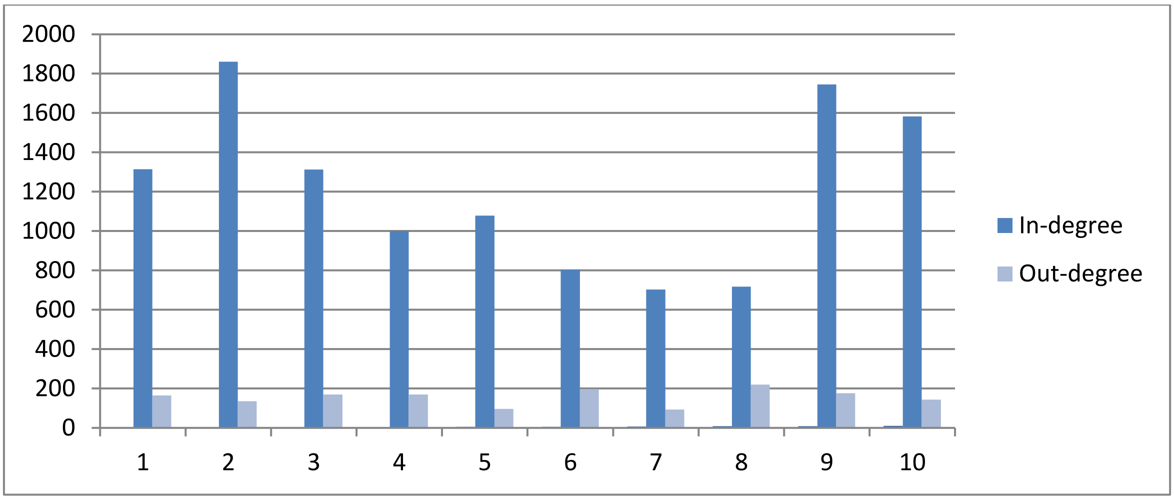

Fig. 6.11 In-Degree and Out-Degree of Overlapping Communities of PCPM

The precision, recall and F-measure of identifying overlapping communities by PCPM is found as 0.77, 0.79 and 0.59 respectively. Table XXXI shows the different analytical measures of PCPM communities.

Table XXXI Analytical Measures of PCPM Communities

Number of Communities	170
Dense Communities	148
Sparse Communities	22
Size of Largest Community	1790
Size of Smallest Community	53
Largest In-Degree	1760
Largest Out-Degree	210
Modularity Score	0.8467654
F measure	0.77
Precision	0.79
Recall	0.59

Comparison of three Clique Percolation Methods

CPM algorithm discovered 198 communities and different size of the node in the graph. OCPM found 180 communities and number of nodes are the dense overlapping community in the network. PCPM exposed 170 communities in the network. It shows better performance than other methods because every overlapping community has large number of nodes in the network. PCPM discovered more number of communities than CPM wherein some communities are sparse and more are overlapping communities. Also, PCPM has shown large modularity score than two algorithms, confirming that there is more interaction between nodes in the network. The comparative results of various critical measures are summarized in Table XXXII.

Table XXXII Different Categories of Communities

Method	Number of Communities	Dense Communities	Sparse Communities	Highest In-Degree	Highest Out-degree	Size of the Largest Community	Size of the Smallest Community	Strong Communities	Weak Communities	Modularity Score
CPM	198	134	64	1698	204	1690	49	123	44	0.7725246
OCPM	180	129	51	1791	214	1710	51	118	34	0.7865626
PCPM	170	148	22	1760	210	1790	53	136	15	0.8467654

The number of iterations in PCPM is less when compared with CPM and OCPM which in turn reduces the computational time of PCPM. In this process, CPM user, system and elapsed time took 83.76, 0.15 and 84.05 respectively. OCPM algorithm’s user, system and elapsed time taken are 37.89, 0.26 and 38.11. Parallel CPM shows better performance than other two methods. It takes less computation time for user, system and elapsed time in the network. Table XXXIII shows user, system and elapsed time for clique percolation method.

Table XXXIII System Elapsed Time

Algorithm	User (seconds)	System (seconds)	Elapsed (seconds)
CPM	83.76	0.15	84.05
OCPM	37.89	0.26	38.11
PCPM	0.29	0.05	0.83

The performance of the three methods CPM, OCPM and PCPM are compared in terms of quality measures precision, recall, F-measure for their efficiency in detecting overlapping communities. The F-measures of identifying overlapping communities by CPM and OCPM are found as 0.73 and 0.75 respectively. Also, PCPM is found to have better recall value of 0.59 when compared to 0.57 recall of CPM. CPM and PCPM algorithm found 0.73 and 0.77 F-measure respectively. PCPM is found to have better precision value of 0.79 when compared to recall of 0.76 by CPM. The comparative results are shown in Table XXXIV and depicted in Fig. 6.14.

Table XXXIV Quality Measure for CPM, OCPM & PCPM

Evaluation Measures	CPM	OCPM	PCPM
F measure	0.73	0.75	0.77
Precision	0.76	0.78	0.79
Recall	0.57	0.58	0.59

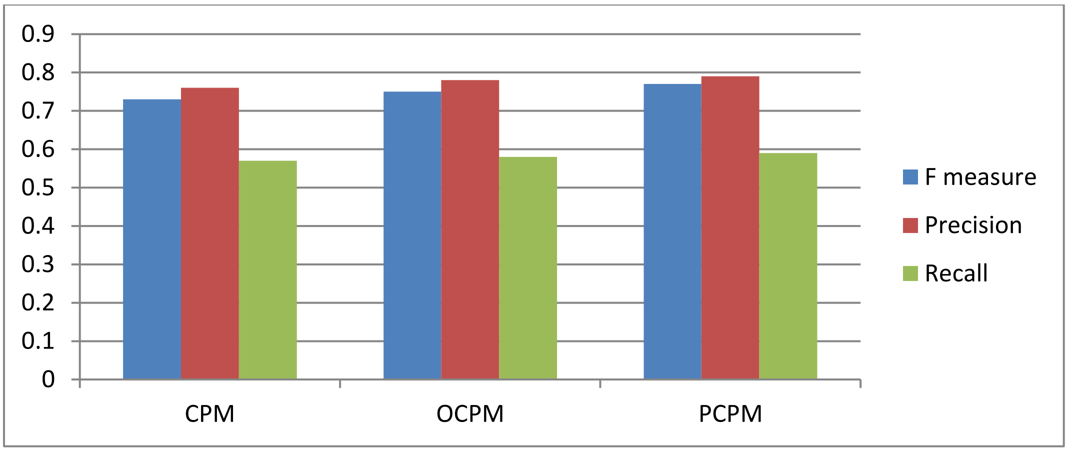

Fig. 6.12 Quality Measure for CPM, OCPM & PCPM

Findings

Three algorithms have been applied for finding overlapping communities in the sports person network in which CPM algorithm discovered more number of communities than OCPM and PCPM. CPM overlapping algorithm discovered 198 communities in the network. OCPM algorithm found 180 different sizes of communities. PCPM algorithm discovered 170 communities and different size of the node in the graph. PCPM yielded better performance compare to other two algorithms as it has used a clique percolation algorithm for detecting clique communities in a network works by inserting its edges and keeping track of the emerging community structure. This algorithm applied on twitter networks, has shown that the computational time required to process a network scales linearly with the number of k-cliques in the network. The modularity score of communities detected by PCPM is large than, that discovered by OCPM and PCPM. The computational time of PCPM algorithm is comparably less than other algorithms CPM and OCPM.

SUMMARY

Clique percolation approach of overlapping community detection has been described and its implementation on sample twitter network data has been elucidated with results and analysis. The two variants of clique percolation based on optimization and parallel processing has been discussed and the comparative analyses were also presented with tables and charts in this chapter. The novel hybrid clique percolation method based on the optimal z-score of k-core communities is designed to identify missing communities and described in the following chapter.