CERTAIN INVESTIGATIONS ON NODE CENTRIC COMMUNITY DETECTION IN SOCIAL NETWORK

Thesis submitted to
BHARATHIAR UNIVERSITY

for the award of the degree of
DOCTOR OF PHILOSOPHY

in
COMPUTER SCIENCE

By
K. SATHIYAKUMARI M.C.A, M.Phil.,

Under the guidance of
Dr. (Mrs.) M. S. VIJAYA M.Sc., M.Phil., Ph.D
Associate Professor and Head
Department of Computer Science (PG)

PSGR Krishnammal College for Women

PSGR KRISHNAMMAL COLLEGE FOR WOMEN
College of Excellence – nirf 22nd Rank
(An Autonomous Institution – Affiliated to Bharathiar University)
Reaccredited with “A” grade by NAAC
An ISO 9001:2015 Certified Institution
Coimbatore – 641 004, Tamilnadu, India

JUNE 2019

8. ENHANCED CLIQUE PERCOLATION FOR OVERLAPPING COMMUNITY DETECTION

Community detection is one of the key gadgets in social network analysis. Detecting the communities involves finding the densely connected nodes. In social networks, nodes can form possibly overlapping communities. Overlapping communities are probable if a node is a member of more than one community. Chapter 6 demonstrated the implementation of clique percolation methods for overlapping community detection wherein the hybrid method of clique percolation experimented on twitter network data has been described in chapter 7. Another approach proposed to enhance the quality of overlapping community detection using association rule mining is presented in this chapter.

8. 1 INTRODUCTION

Detecting communities in networks is one of the most popular topics of network science. The existence of community structure indicates that the nodes of the network are not homogeneous but divided into classes, with a higher probability of connections between nodes of the same class than between nodes of different classes. Many approaches to community detection exist, spanning not only different algorithms and partitioning strategies but also with fundamentally different definitions of a community.

Community detection in a social network is a prominent issue in the study of the network system as it helps to understand the structure of a network. A member of a social network can be part of more than one group or community. As a member can be ovearrlapped between more than one groups, specific overlapping community detection techniques are essential in order to identify the overlapping nodes. One of the first algorithms allowing shared members between the communities is given by the clique percolation method. In this, the basic building blocks of the communities are given by k-cliques and communities are associated with k-clique percolation clusters. The usual rule for finding the optimal partitioning in this approach is to tune the system to the critical point of k-clique percolation. The reason behind this rule is the emergence of a giant percolating community by merging many smaller communities, thereby finding a community structure as highly structured as possible. The problem with this method is that it does not cover the complete network. Hence some nodes may not be a part of any community irrespective of their connectivity. In this work, a novel approach has been introduced to extend the clique percolation method so that each and every connected node will be part of at least one community.

Association rule mining is a procedure which aims to observe frequently occurring patterns, correlations, or associations found in various kinds of databases. In social networks, where the users carry out their preferred activities, the actions of users on networks are perceived as patterns from which semantic and significant associations between the users can be drawn. Hence, it is intended to discover frequent patters and association rules based on the similarity in user's interests and activities of users on networks. These rules will support the clique percolation method to cover the missing nodes and overlapping communities while community detection [95].

8.2 ENHANCED CLIQUE PERCOLATION METHOD

Association rule learning is a prominent and a well-explored method for determining relations among variables in large databases. Apriori algorithm is commonly used association rule learning for mining frequent item sets and relevant association rules. In social networks, the interaction between the nodes is conceived as patterns and the association rule mining is used to generate rules from twitter dataset based on frequents occurrence of similar interest nodes to help the decision of communities.

Each pattern includes a sequence of users having a similar pattern in the network. Each user sequence is a homogeneous group with minimum support of frequent pattern mining. Users in each homogeneous group reveal node with similar nodes performance and interest in the network, and node in these groups can act as members of the group if they are linked to each other. This link does not only mean a direct connection or the existence of edges among nodes, but also mean communication with mediator nodes accepted in a threshold. The outputs from this process are a group of users, in which similar and related users are its members. Such groups are called small communities. Each small community is viewed as a core of one community, and neighboring people are considered followers of these small communities. Thus, small communities are expanded by taking into consideration communication and changes into acceptable community sizes.

The network is assumed as set E including n nodes, E = {e_1, e_2, e_3... e_n}, where each node e_j is linked to the dataset T_j that represents the edges relating to this node. Thus, for each node, there is one transaction data set. Using closed frequent patterns, this set of data can be converted into a vector, which is considered as an effective representative of edges related node. Apriori algorithm is used to mine the rules from T_j. based on the number of matching items greatly and improve the efficiency of community detection [96].

The association rules generated will help in determining the connected components in the network G efficiently and searching the adjacent cliques of size k in clique percolation. The ECPM can discover all possible overlapping cliques from the network in K+1iterations.

Algorithm

Input: Graph G, clique size k

Output: Overlapping Communities C

Process:

The network, G and the clique size, k

Step 1: D ← Generate data from twitter

Step 2: Read data using function X= read. Transactions (D, Sep=” “)

Step 3: Define rules with “rules=@user name”, using apriori algorithm

Step 3: Calculate P=apriori(X, parameter = list (support = 0.006, confidence = 0.25, minlen = 2))

Step 4: Generate the rules

Step 5: Discover overlapping communities using CPM-clique (P, k)

8.3 ENHANCED OVERLAPPING COMMUNITY DETECTION MODEL

The three important elements of the proposed method are: (i) construction of network (ii) association rule-based clique percolation (iii) overlapping communities and quality measures. The twitter data of a sports person’s network drawn using Twitter API and raw data converted to edgelist is used in the first phase. The second element portrays the core process which involves identification of frequent patterns and generation of association rules to discover initial communities with similar interests and clique percolation to locate overlapping communities. Eventually, the overlapping communities and the corresponding quality measures are found and the performance enhancement is demonstrated in the third segment. The architecture of the enhanced overlapping community detection model is shown in Fig. 8.1.

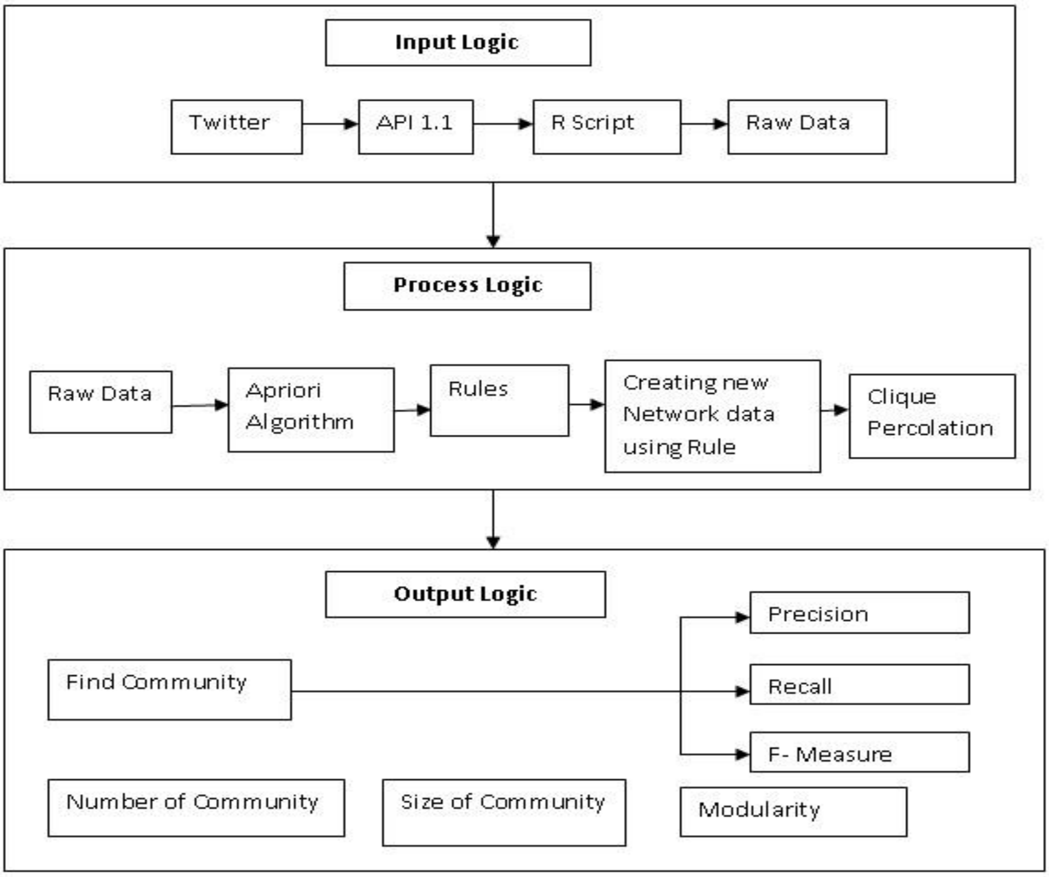

Fig. 8.1 Enhanced Overlapping Community Detection Model

8.4 EXPERIMENTS AND RESULTS

The experiment is carried out first by extracting association rules using the raw data of the given twitter network. Apriori algorithm is used and implemented through R scripts to generate rules. The frequent pattern output is a sequence of nodes that are found in a similar interest group. The ECPM overlapping community detection algorithm based on association rules of frequent patterns attempts to identify the similar interest groups from the twitter network. The sample association rules discovered from the twitter data is given below.

```
"rules='@imVkohli","support=90","confidence","friends/followes","count"
"{444} => {377}",0.00649350649350649,0.666666666666667,47.3846153846154,6
"{377} => {444}",0.00649350649350649,0.461538461538462,47.3846153846154,6
"{444} => {339}",0.00649350649350649,0.666666666666667,38.5,6
"{339} => {444}",0.00649350649350649,0.375,38.5,6
"{333} => {360}",0.00649350649350649,0.857142857142857,27.3103448275862,6
"{249} => {175}",0.00649350649350649,0.857142857142857,79.2,6
"{175} => {249}",0.00649350649350649,0.6,79.2,6
"{377} => {339}",0.00649350649350649,0.461538461538462,26.6538461538462,6
"{339} => {377}",0.00649350649350649,0.375,26.6538461538462,6
```

```
"{259} => {260}",0.00649350649350649,1,132,6
"{260} => {259}",0.00649350649350649,0.857142857142857,132,6
"{260} => {321}",0.00649350649350649,0.857142857142857,60.9230769230769,6
"{321} => {260}",0.00649350649350649,0.461538461538462,60.9230769230769,6
"{266} => {472}",0.00649350649350649,0.666666666666667,30.8,6
"{472} => {266}",0.00649350649350649,0.3,30.8,6
"{250} => {438}",0.00649350649350649,0.666666666666667,25.6666666666667,6
"{438} => {250}",0.00649350649350649,0.25,25.6666666666667,6
"{250} => {360}",0.00649350649350649,0.666666666666667,21.2413793103448,6
"{366} => {381}",0.00757575757575758,0.7,43.12,7
"{381} => {366}",0.00757575757575758,0.466666666666667,43.12,7
"{366} => {221}",0.00757575757575758,0.7,64.68,7
"{221} => {366}",0.00757575757575758,0.7,64.68,7
"{474} => {415}",0.00757575757575758,0.636363636363636,23.52,7
"{415} => {474}",0.00757575757575758,0.28,23.52,7
"{381} => {221}",0.00649350649350649,0.4,36.96,6
"{221} => {381}",0.00649350649350649,0.6,36.96,6
"{381} => {472}",0.00649350649350649,0.4,18.48,6
"{472} => {381}",0.00649350649350649,0.3,18.48,6
"{381} => {454}",0.00974025974025974,0.6,15.84,9
"{454} => {381}",0.00974025974025974,0.257142857142857,15.84,9
"{381} => {360}",0.00649350649350649,0.4,12.7448275862069,6
"{135} => {160}",0.00649350649350649,0.666666666666667,38.5,6
"{160} => {135}",0.00649350649350649,0.375,38.5,6
"{76} => {160}",0.00865800865800866,1,57.75,8
"{160} => {76}",0.00865800865800866,0.5,57.75,8
"{476} => {450}",0.00649350649350649,0.375,19.25,6
"{450} => {476}",0.00649350649350649,0.333333333333333,19.25,6
"{476} => {417}",0.00757575757575758,0.4375,14.9722222222222,7
"{417} => {476}",0.00757575757575758,0.259259259259259,14.9722222222222,7
"{320} => {160}",0.00649350649350649,0.428571428571429,24.75,6
"{160} => {320}",0.00649350649350649,0.375,24.75,6
"{426} => {417}",0.00649350649350649,0.25,8.55555555555556,6
"{325} => {407}",0.00649350649350649,0.333333333333333,19.25,6
"{407} => {325}",0.00649350649350649,0.375,19.25,6
"{325} => {417}",0.00757575757575758,0.388888888888889,13.3086419753086,7
"{417} => {325}",0.00757575757575758,0.259259259259259,13.3086419753086,7
"{325} => {454}",0.00649350649350649,0.333333333333333,8.8,6
"{325} => {360}",0.00649350649350649,0.333333333333333,10.6206896551724,6
"{472} => {454}",0.00865800865800866,0.4,10.56,8
"{244} => {160}",0.00649350649350649,0.375,21.65625,6
"{160} => {244}",0.00649350649350649,0.375,21.65625,6
"{182} => {310}",0.00649350649350649,0.375,21.65625,6
"{310} => {182}",0.00649350649350649,0.375,21.65625,6
```

The association rules are then converted into the format as required by CPM and is given as input edgelist to the ECPM. This approach discovered 181communities in the network out of which ninety-nine communities have 150 members in the community network. 99 communities are having the large number of nodes accounting to 1800 to 501 sizes of the nodes in the community. Seventy communities are having the medium number of nodes that is 500 to 101 in the community. Twelve communities are having the small number of nodes and community sizes ranging from 20 to100 in the network. The ECPM algorithm found 181 dense communities in the network, which depicts this network has large number of nodes and shares the link for each node. Fig. 8.2 shows overlapping communities detected from cricket player’s network by ECPM.

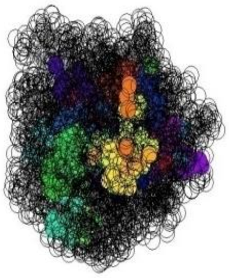

Fig. 8.2 Communities Detected By ECPM

A sample of 10 communities detected by ECPM with node ids is given below.

```
[14]   21 77 714 701 699 697 696 694 693 692 690 689 687 686 684 683 681 680 679
[15]  20  42 702 701 699 697 696 694 693 692 690 689 687 686 684 683 681 680 679 678 381
[16]  19  99 148 675 674 672 668 667 666 663 662 660 654
[17]  17  18 615 614 613701 699 697 696 694 693 692 690 689 687 686 684 683 681 680 679  612 605 331 220
[18]   14 95 535 534 533 527 523 510 508 701 699 697 696 694 693 692 690 689 687 686 684 683 681 680 679
[19]  13  27 502 501 499 495 492 260 701 699 697 696 694 693 692 690 689 687 686 684 683 681 680 679 335
66   92 1817 1814 1803 73  104 1960 1959 1956 1955 1954 1953 1951 1950 1949 1947 1946 1945 1944 1658
1356 1308
[20]  6  67 305 303 296 289 283 282 1191 1190 1188 1187 1186 1185 1184 1180 1176 1111
[21]  5  34 272 269 266 262 254 252 249 244 155 1191 1190 1188 1187 1186 1185 1184 1180 1176 1111
[22]   3 37 206 114 202 200 198 196 194 193 189 186 185 182 181 1191 1190 1188 1187 1186 1185 1184 1180
1176 1111 335    66   92 1817 1814 1803 73 104 1960 1959 1956 1955 1954 1953 1951 1950 1949 1947 1946
1945 1944 1658 1356 1308
[23] 2 87 179 177 175 171 170 168 166 165 159 156 154 1191 1190 1188 1187 1186 1185 1184 1180 1176
1111 335    66   92 1817 1814 1803 73 104 1960 1959 1956 1955 1954 1953 1951 1950 1949 1947 1946 1945
1944 1658 1356 1308
[24]   1   9 146 145 142 141 140 138 137 131 130 129 124 123 122 118 117 335    66   92 1817 1814 1803 73
104 1960 1959 1956 1955 1954 1953 1951 1950 1949 1947 1946 1945 1944 1658 1356 1308
[25] 22 31 744 335    66   92 1817 1814 1803 73 104 1960 1959 1956 1955 1954 1953 1951 1950 1949 1947
1946 1945 1944 1658 1356 1308
```

The modularity score obtained is 0.87. 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 181, there are 169 dense communities and 12 sparse communities detected. The size of the largest community obtained is 1750 and the size of the smallest subgroup is 57. The membership distribution of nodes for a sample of 10 communities is given in Table XXXXII and illustrated in Fig. 8.3.

Table XXXXII Sizes of ECPM Overlapping Communities

Overlapping Communities	Size of Communities
ECPM 1	1370
ECPM 2	1280
ECPM 3	1690
ECPM 4	1003
ECPM 5	1743
ECPM 6	1010
ECPM 7	1570
ECPM 8	1769
ECPM 9	1632
ECPM 10	1303

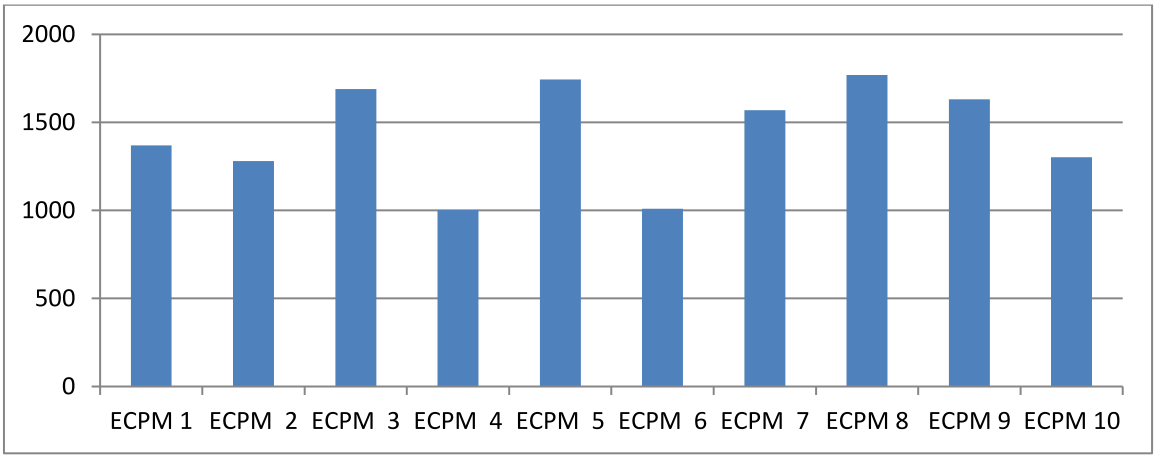

Fig. 8.3 Sample Communities and their Sizes Identified by ECPM

Also, it is observed from the empirical results that that in-degree of 91communities lies between 501 to 1800 and the in-degree of 70 communities lies between 101 to 500 which indicate that friends and followers are more interactive with other nodes. The in-degree of 20 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 116 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 65 communities, the out-degree lies in the range of 20 to 60. The degrees measures evaluated using ECPM and the results for samples of 10 communities are presented in Table XXXXIII and illustrated in Fig. 8.4.

Table XXXXIII Degree Measures of ECPM Communities

ECPM Communities	1	2	3	4	5	6	7	8	9	10
In-degree	1405	968	983	789	974	881	736	678	1566	983
Out-degree	80	78	96	101	74	86	91	112	127	95

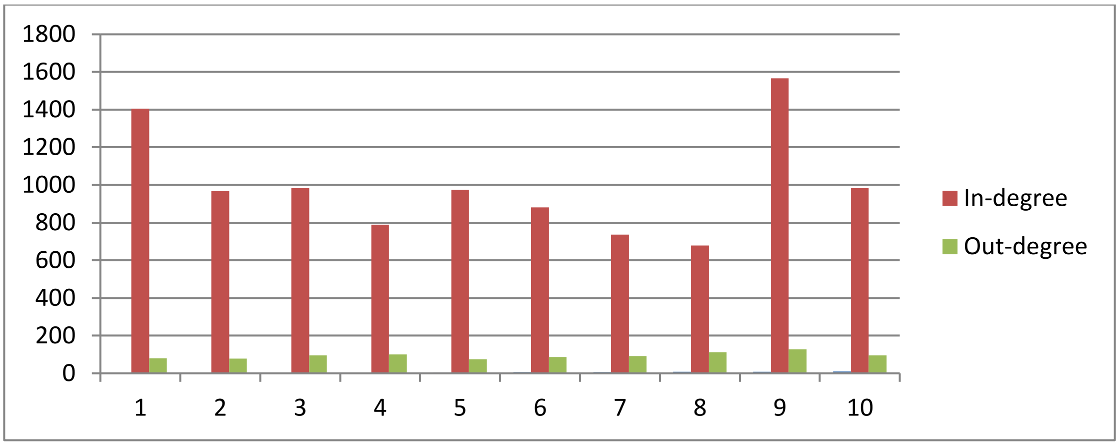

Fig. 8.4 In-Degree and Out-Degree of 10 ECPM Overlapping Communities

The effectiveness of ECPM in identifying missing nodes and overlapping communities is determined using measures like precision, recall, F-score by comparing the predicted communities against the ground truth communities of the given network as described in section 6.4. The ECPM yielded the results of precision as 0.82 whereas the recall and the F-measure is 0.7and 0.81 respectively. The results of various analytical measures are tabulated in Table XXXXIV.

Table XXXXIV Analytical Measures of ECPM Communities

Number of Communities	181
Dense Communities	169
Sparse Communities	12
Size of Largest Community	1750
Size of Smallest Community	57
Largest In-Degree	1699
Largest Out-Degree	207
Modularity Score	0.87
F measure	0.81
Precision	0.82
Recall	0.7

Comparison of ECPM, HCPM, PCPM, OCPM and CPM

The effectiveness of enhanced clique percolation method is compared with basic three CPMs and the hybrid CPM approach presented in chapter 7. The CPM algorithm discovered 198 communities and different size of the node in the graph. In this network, OCPM, PCPM algorithm discovered 180 and 170 communities respectively. HCPM found 220 communities and number of nodes are the dense overlapping community in the network. The numbers of nodes were dense in the overlapping community. HCPM method has found more number of communities than CPM and ECPM exposed 181 communities in the network. It showed better performance than other methods because every overlapping community has large number of nodes in the network. ECPM discovered more number of communities than CPM wherein some communities are sparse and more are overlapping communities. Also, ECPM has shown large modularity score than basic three CPMs, confirming that there is more interaction between nodes in the network. ECPM has outperformed other two methods in detecting sparse communities. The comparative results of various critical measures are summarized in Table XXXXV.

Also, the ECPM modularity score of 0.87 is much higher than 0.77 achieved by CPM. OCPM and PCPM algorithm found 0.78 and 0.84 modularity respectively. Networks detected by ECPM have high modularity and so dense connections exist between the nodes within modules but sparse connections between nodes in different modules.

ECPM discovered the largest community of size 1794 nodes when compared to CPM which found a community of strength 1690. Moreover, ECPM discovered more number of communities than CPM wherein some communities are sparse and more are overlapping communities and ECPM yields more number of good communities than CPM and HCPM.

Table XXXXV 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
HCPM	220	172	48	1661	213	1610	59	148	58	0.85
ECPM	181	169	12	1695	207	1794	57	141	9	0.87

As the aim of ECPM is to cover all the overlapping nodes, time complexity of ECPM is not taken into account for comparison with three basic CPMs.

It was shown that with increasing flexibility the quality of the community enhances with the proposed methods with regard to the evaluation metrics *F* measure, precision and recall. CPM and ECPM algorithm found F-measure of 0.73 and 0.81 respectively. The F-measures of identifying overlapping communities by OCPM and PCPM are found as 0.75 and 0.77 respectively. PCPM is found to have better precision value of 0.79 when compared to precision of 0.78 by OCPM. The F-measures of identifying overlapping communities by CPM and HCPM are found as 0.73 and 0.78 respectively. Also, HCPM is found to have better recall value of 0.69 when compared to 0.57 recall of CPM. ECPM is found to have better precision value of 0.82 when compared to precision of 0.76 by CPM. The performance evaluation of all the five methods with respect to quality measures such as precision, recall, F-measure obtained against ground truth communities are shown in Table XXXXVI and illustrated in Fig. 8.5.

Table XXXXVI Quality Measure for CPM, OCPM, PCPM, HCPM & ECPM

Measures	**CPM**	**OCPM**	**PCPM**	**HCPM**	**ECPM**
Precision	0.76	0.78	0.79	0.79	0.82
Recall	0.57	0.58	0.59	0.69	0.7
F measure	0.73	0.75	0.77	0.78	0.81

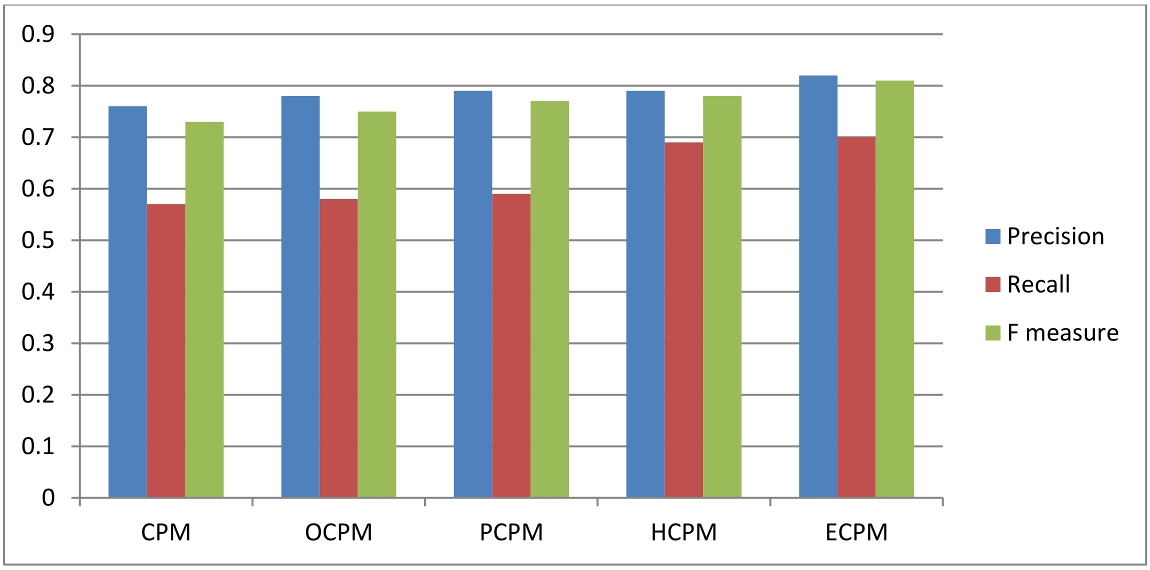

Fig. 8.5 Quality Measure for CPM, OCPM, PCPM, HCPM & ECPM

Findings

From the comparative results, it is found that various analytical measures produced by ECPM are improved than CPM and its variants. The modularity score of communities detected by ECPM is better than that of PCPM and larger than that discovered by CPM and OCPM. The high modularity of ECPM confirms a higher density of communities. ECPM is effective than CPM, OCPM and PCPM in identifying the missing nodes as the cliques are discovered based on association rules of user's similar interests. The proposed enhanced clique percolation method outperforms in recognizing overlapping communities than CPM, OCPM, PCPM as the evaluation metrics precision, recall, and F-measure are high in ECPM. The empirical result analysis of ECPM algorithm on twitter network data and exhaustive experiments of various overlapping community detection algorithms described in chapter 6 and chapter 7, ascertain that the community detection quality has been enhanced with association rule mining of frequent networking group.

SUMMARY

Clique percolation method of overlapping community detection has been enhanced using frequent patterns of similar interest groups and its demonstration on sports person's network data has been elucidated in this chapter with results and analysis. The quality of community detection and the effectiveness of enhanced clique percolation method in detecting missing nodes and overlapping communities evaluated using ground truth communities is analyzed with various performance metrics. The comparative analyses of ECPM with three basic CPMS were also presented with tables and charts in this chapter. Overall the present work demonstrating the enhanced method principals to improve the community detection quality and will bring more meticulousness in the network community detection.

9. CONCLUSION

The thesis titled “Certain Investigations on Node-Centric Community Detection in Social Network” portrays the research work carried out on community detection analysis of twitter network data using social network mining.

Social network analysis and social network mining has been carried out by modeling the twitter network data into a graph. As twitter networks are directed networks, node centric approach has been employed in this research to investigate various community detection approaches. Basic community detection, sub-community detection and overlapping community detection have been implemented using graph partitioning techniques. Two hybrid approaches have been proposed for efficient overlapping community detection.

A real-time twitter network data has been crawled from Twitter using R3.5.1 at run time. The crawled data consists of 19000 nodes of friends and followers list of a sports person. This sample network data was used throughout the research to carry out various investigations on community detection.

The work has been carried out in five stages. In the first stage, social network analysis of the sample twitter network has been done by modeling the network data into a graph structure and analyzed using graph properties like degree, closeness, betweenness. The modularity score of 0.91 proved that the sports person’s friends and followers network is highly dense. Next, the graph-partitioning algorithm based community detections are performed by dividing the nodes of a network into groups using Girvan-Newman edge betweenness and random walk algorithm for discovering fundamental communities. In the third stage, maximum-k clique, maximum-k core, maximum-k plex based subgraph analysis has been carried out for sub-community detection and the performance was analyzed.

Ground truth communities were created manually based on the structural properties of network and the results of k-plex and maximal clique. Subsequently, clique percolation and its variants such as optimized clique percolation method and parallel clique percolation method have been implemented to discover overlapping communities. The performance of these methods was evaluated by comparing the predicted communities against ground truth communities using precision, recall, F-measure.

Finally, two hybrid approaches designed using optimal z score of k-core and using association rule mining of frequent patterns, have been implemented for discovering overlapping communities using the same network data. The performance of these models was evaluated and comparative analysis was done with respect to the evaluation metrics such as precision, recall, F-measure. The observations made and the interpretations drawn from this research work are summarized below.

- The sample network used in this research for investigations on community detection has high degree centrality which shows that the node is active and having advantaged position in the network. The low closeness centrality depicts that the node has slow interaction to other entities in a network. Also, the node is in a powerful position and a better influence over the other nodes in the network because the betweenness centrality is high for this network. The modularity score of 0.91 proved that the sports person's friends and followers network is highly dense.
- Principal communities discovered through graph partitioning illustrated that the Girvan-Newman algorithm has detected more number of communities and few communities are dense. The random walks detected less number of communities and all communities are very dense. The modularity score showed by Girvan-Newman is better than that of random walks.
- The maximal k-core algorithm detected subgroups based on the k-core value from the twitter network. The k-core size of 3 delivered more number of dense communities and sparse communities. Also the size of the sparse community is less in case of k-core. Therefore, the k-core algorithm depicts higher communication between the nodes.
- The k-plex establishes the intractability of the communities for every fixed k as it is a graph-theoretic relaxation of cliques and confirms higher interaction between friends and followers.
- The maximal k-clique shows more number of strong communities as the degree of the communities detected by k-clique is higher than maximal k-core and k-plex.
- The maximal k-core algorithm yielded high modularity score which ascertains the better community detection quality. Among all three subgraph algorithms, k-core establishes superiority community detection measures.
- CPM algorithm discovered more number of communities when compared to OCPM and PCPM. The modularity score of communities detected by PCPM is large than, that discovered by CPM and OCPM. The computational time of PCPM algorithm is comparably less than other algorithms CPM and OCPM.

- The hybrid clique percolation approach performs better than CPM in recognizing overlapping communities as the evaluation metrics such as *F* measure, precision, recall is high. Hybrid clique percolation has recognized more number of dense communities and less number of sparse communities than CPM.
- Experiment on overlapping community detection proved to be effective in identifying the the small, medium and big communities based on the size and the interaction between the nodes.
- The performance of clique percolation is enhanced with frequent pattern mining in recognizing overlapping communities and confirmed as its scores precision; recall and F-measure are high. Also, the higher modularity score of the network communities produced by ECPM proved higher density of communities.

The research contributions made in thesis are listed below.

- Real time network data was created from the Twitter account of a sports person using Twitter API 1.1.
- Implemented graph clustering algorithms on directed network in node centric based community detection for detecting principal communities
- Implemented sub graph analysis for sub community detection using maximum k- clique, maximum k- core, maximum k-plex techniques
- Ground truth communities are defined and labeled manually based on the results of k-plex and maximal clique with the support of measures like in-degree, out-degree, in-closeness, out-closeness and betweenness
- Implemented overlapping community detection methods on twitter data using clique percolation and its variants such as optimization and parallel clique percolation
- Proposed a hybrid overlapping community detection approach using z-score and k-core based clique percolation
- Enhanced the performance of clique percolation using association rule mining for overlapping community detection

The community detection problem is one of the challenging tasks in social network analysis. This research work demonstrated different models for principal community, sub community and overlapping community detections and examined through twitter network data of a sport person. The comprehensive research work was carried out using three notions of graph theory such as graph clustering, sub-graph analysis, clique percolation. The exhaustive experiments performed on various models showed significant improvement in community detection quality with respect to performance metrics.

As the scope for future work the community detection algorithms can be employed on weighted directed graph. Since the social network graph is large, map-reduce programming model for big data can be adopted. The problem can be extended for various community detection methods like group, network and hierarchy centric community.