CONTENTS

Acknowledgement iv

List of Tables v

List of Figures vii

List of Abbreviations x

Abstract xii

1. INTRODUCTION 1

1.1 Data Mining and Social Network Analysis 1

1.2 Data Mining in Social Network 12

1.3 Community Detection Approaches 17

1.4 Review of Literature 20

1.5 Objectives of the Research 28

1.6 Organization of the Thesis 29

2. COMMUNITY DETECTION IN SOCIAL NETWORK 31

2.1 Graph Theory - Definition and Terminology 31

2.2 Community and its Types 35

2.3 Community Detection Approaches 39

2.4 Node-Centric Community Detection Algorithms 42

3. SAMPLE TWITTER NETWORK AND ITS ANALYSIS 47

3.1 Introduction 47

3.2 Twitter Network Data 48

3.3 Network Structure of Twitter Network Data 54

3.4 Network Analysis of Sample Twitter Data 58

4. DIRECTED NETWORK BASED ON COMMUNITY DETECTION 66

4.1 Introduction 66

4.2 Girvan-Newman Edge Betweenness Algorithm 68

4.3 Random Walks Algorithm 70

4.4 Directed Network Community Detection Model 73

4.5 Experiments and Results 74

5. COMMUNITY DETECTION BASED ON SUBGRAPH ANALYSIS 79

5.1 Introduction 79

5.2 Subgraph Algorithms 80

5.2.1 Maximal K-Clique Algorithm 80

5.2.2 Maximal K-Core Algorithm 82

5.2.3 Maximal K-Plex Algorithm 84

5.3 Subgraph Based Community Detection Model 86

5.4 Experiments and Results 87

6. OVERLAPPING COMMUNITY DETECTION USING CLIQUE PERCOLATION 101

6.1 Introduction 101

6.2 Overlapping Community Detection Algorithm 102

6.3 Overlapping Community Detection Model 106

6.4 Ground Truth Communities and Evaluation 107

6.5 Experiments and Results 110

7. NOVEL HYBRID CLIQUE PERCOLATION ON COMMUNITY DETECTION 124

7.1 Introduction 124

7.2 Hybrid Clique Percolation Method 125

7.3 Hybrid Overlapping Community Detection Model 127

7.4 Experiments and Results 128

8. ENHANCED CLIQUE PERCOLATION ON COMMUNITY DETECTION 135

8.1 Introduction 135

8.2 Enhanced Clique Percolation Method 136

8.3 Enhanced Overlapping Community Detection Model 137

8.4 Experiments and Results 138

9. CONCLUSION 146

REFERENCES 150

List of Publications 158

Appendix

A. Datasets 159

B. Coding 170

C. Screenshots 189

ACKNOWLEDGMENT

The outset, I would like to express my humble thanks to **The God Almighty**, for the kind grace showered on me to complete the research work successfully.

I express my sincere thanks to **Smt. R. Nandini,** Chair Person, PSGR Krishnammal College for Women, Coimbatore for giving me this opportunity to undergo Ph.D. in Computer Science in their esteemed institution.

I express my whole-hearted thanks to **Dr. (Mrs.) N. Yesodha Devi, M.Com., Ph.D,** Secretary, PSGR Krishnammal College for Women, Coimbatore for her constant encouragement during my course of study.

I thank **Dr. (Mrs.) S. Nirmala., MBA., M.Phil., Ph.D,** Principal, PSGR Krishnammal College for Women, Coimbatore for the support during my course of study.

I wish to place on record my deep sense of gratitude to my research supervisor **Dr. (Mrs). M. S. Vijaya, M.Sc., M.Phil., Ph.D,** Associate Professor and Head, Department of Computer Science, Coimbatore for suggesting this field of research and for her constant guidance, motivation and untiring help during study. I am indebted to her for the kindness and patience she exhibited throughout the course of study which enabled me to successfully complete my research work.

I am also very grateful to my friends and research colleagues **Dr. S. Karpagavalli** and **Mrs. V. Pream Sudha** for their insightful discussions and suggestions. Their encouragement also helped me to overcome the difficulties encountered in my research.

I am obliged to all my colleagues for the support extended throughout my study.

I am very much indebted to all my family members for their invaluable moral support rendered during my course of study. Finally, I would like to thank all the people who have supported me for the successful completion of my research work.

K. Sathiyakumari

LIST OF TABLES

TABLE No.	TITLE	PAGE No.
Table I	Blogs of Social Media	12
Table II	Summary of Literature Survey	26
Table III	Symbols and Definitions	31
Table IV	Sample Twitter Data of Sports Person with Friends and Followers	53
Table V	Adjacency Matrix of Twitter Data	56
Table VI	Sample Nodelist of Twitter Data	57
Table VII	Sample Edgelist of Twitter Data	58
Table VIII	Measures of Sample Twitter Network Data	63
Table IX	Sample Value of Eigen Centrality	64
Table X	Communities Identified by Edge betweenness and Respective Sizes	75
Table XI	Communities Identified by Random Walks and Respective Sizes	77
Table XII	Community Analysis for Edge-Betweenness and Random Walks	77
Table XIII	Community Matrix of Adjacency Matrix	88
Table XIV	K-Clique Subgroups When k=3	89
Table XV	Centrality Measures of Communities	89
Table XVI	Size of K-Core Subgroups for K= 1 To 5	91
Table XVII	Degree Measures of K-Core Subgraphs	94
Table XVIII	Size of K-Core Sub-Groups when k=3	94
Table XIX	In-Degree and Out-Degree of K-Core Subgroups when k=3	95
Table XX	Size of K-Plex Sub-Groups when k=3	97
Table XXI	In-Degree and Out-Degree of the K-Plex Subgroups	98
Table XXII	Subgraph Analysis for K-Clique, K-Core and K-Plex	99
Table XXIII	Sizes of CPM Overlapping Communities	112
Table XXIV	Degree Measures of CPM Communities	113
Table XXV	Analytical Measures of CPM Communities	113

Table XXVI	Sizes of OCPM Overlapping Communities	115
Table XXVII	Degree Measures of OCPM Communities	116
Table XXVIII	Analytical Measures of OCPM Communities	117
Table XXIX	Sizes of PCPM Overlapping Communities	118
Table XXX	Degree Measures of PCPM Communities	119
Table XXXI	Analytical Measures of PCPM Communities	120
Table XXXII	Different Categories of Communities	121
Table XXXIII	System Elapsed Time	121
Table XXXIV	Quality Measure for CPM, OCPM & PCPM	122
Table XXXV	K-Core Size of Sub-Groups	128
Table XXXVI	Sizes of HCPM Overlapping Communities	130
Table XXXVII	Degree Measures of HCPM Communities	131
Table XXXVIII	Analytical Measures of HCPM Communities	131
Table XXXIX	Comparative Results of HCPM with other Three CPMs	132
Table XXXX	System Elapsed Time	132
Table XXXXI	Quality Measure of HCPM and Three CPMs	133
Table XXXXII	Sizes of ECPM Overlapping Communities	141
Table XXXXIII	Degree Measures of ECPM Communities	142
Table XXXXIV	Analytical Measures of ECPM Communities	142
Table XXXXV	Different Categories of Communities	143
Table XXXXVI	Quality Measure for CPM, OCPM, PCPM, HCPM & ECPM	144

LIST OF FIGURES

FIGURE No.	TITLE	PAGE No.
Fig. 1.1	Social Network Showing Nodes and Links	10
Fig. 1.2	Social Network Community Structure	14
Fig. 2.1	A Sample Graph	32
Fig. 2.2	A Directed Graph and Undirected Graph	33
Fig. 2.3	Disjoint Communities	37
Fig. 2.4	Overlapping Communities	38
Fig. 3.1	Twitter Network Data Crawling Process	50
Fig. 3.2	Getting Access Token and Secret Keys in Twitter API	51
Fig. 3.3	Generation of Seven Digit Identification Number in Twitter API	51
Fig. 3.4	Authorization and Virtual Connection to Twitter	52
Fig. 3.5	Sports Person's Network with Friends and Followers	55
Fig. 3.6	Relationship between Users with Friends and Followers	56
Fig. 3.7a	In-Degree of Given Network	62
Fig. 3.7b	Out-Degree of Given Network	62
Fig. 3.8a	In-Closeness of Given Network	63
Fig. 3.8b	Out-Closeness of Given Network	63
Fig. 4.1	Number of Shortest Paths in a Graph	69
Fig. 4.2	Community Detection Framework	74
Fig. 4.3	Communities Detected using Edge Betweenness Algorithm	75
Fig. 4.4	Edge-Betweenness Community Size	75
Fig. 4.5	Communities Detected using Random Walk Algorithm	76
Fig. 4.6	Random Walk Community Size	76
Fig. 4.7	Comparative Results of Edge-Betweenness and Random Walks	78
Fig. 5.1	Cliques with 1, 2, 3, 4, 5 and 6 Vertices	81
Fig. 5.2	K-Core for Small Graph	83

Fig. 5.3	Illustration of k-plexes for k = 1, 2, 3	85
Fig. 5.4	Sub Graph Community Detection Framework	87
Fig. 5.5	Maximal Clique Subgroups of the Twitter Network	88
Fig. 5.6	1-12 K-Core Sub Groups	90
Fig. 5.7	Community Distribution of K-Core Subgraphs	91
Fig. 5.8	Maximal K- Core Subgraphs when k=1	92
Fig. 5.9	Maximal K- Core Subgraphs when k=2	92
Fig. 5.10	Maximal K- Core Subgraphs when k=3	92
Fig. 5.11	Maximal K- Core Subgraphs when k=4	93
Fig. 5.12	Maximal K- Core Subgraphs when k=5	93
Fig. 5.13	In-Degree and Out-Degree of Maximal K- Core Sub Graphs when k=3	95
Fig. 5.14a	Total K-plex Subgroup Network	96
Fig. 5.14b	K-Plex Sub Graph Total Network	96
Fig. 5.15	In-Degree and Out-Degree of Maximal K- Plex Sub Graphs	98
Fig. 6.1	Illustration of Overlapping Community Detection by CPM	103
Fig. 6.2	Overlapping Community Detection Model	107
Fig. 6.3	Communities detected of CPM	111
Fig. 6.4	Size of CPM Overlapping Community	112
Fig. 6.5	In-Degree and Out-Degree of Overlapping Communities by CPM	113
Fig. 6.6	OCPM Community Size and Network	114
Fig. 6.7	Sizes of OCPM Overlapping Communities	115
Fig. 6.8	In-Degree and Out-Degree of Overlapping Community OCPM	116
Fig. 6.9	PCPM Overlapping Community Network	117
Fig. 6.10	Sizes of PCPM Overlapping Community	119
Fig. 6.11	In-Degree and Out-Degree of Overlapping Communities of PCPM	120
Fig. 6.12	Quality Measure for CPM, OCPM & PCPM	122
Fig. 7.1	Proposed Clique Percolation Framework	127

Fig. 7.2	Communities Detected By HCPM	129
Fig. 7.3	Sizes of HCPM Overlapping Community	130
Fig. 7.4	In-Degree and Out-Degree of Overlapping Communities by HCPM	131
Fig. 7.5	System Elapsed Time	133
Fig. 7.6	Comparison of HCPM Quality Measures with CPM, OCPM, PCPM	133
Fig. 8.1	Enhanced Overlapping Community Detection Model	138
Fig. 8.2	Communities Detected By ECPM	140
Fig. 8.3	Sample Communities and Their Sizes Identified By ECPM	141
Fig. 8.4	In-Degree and Out-Degree of 10 ECPM Overlapping Communities	142
Fig. 8.5	Quality Measure for CPM, OCPM, PCPM, HCPM & ECPM	144