Community Detection in Social Networks Using Improvised Evolutionary Algorithms

Thesis Submitted for the award of Degree of

Doctor of Philosophy

In Computer Science & Engineering

Supervisor

Dr. Bhaskar Biswas
(Associate Professor)

By

Harish Kumar Shakya
(Research Scholar)
Er No- 341685
Roll No. 11600EN003

Department of Computer Science & Engineering
Indian Institute of Technology
(Banaras Hindu University)
Varanasi-221005
INDIA

2018

Department of Computer Science & Engineering
Indian Institute of Technology
(Banaras Hindu University)
Varanasi – 221 005

Undertaking from the Candidate

*I, **Harish Kumar Shakya**, declare that the work embodied in this Ph.D. thesis is my own bonafide work carried out by me under the supervision of **Dr. Bhaskar Biswas** for a period of about 06 years, 06 Months from July, 2011 to January, 2018 at the Department of Computer Science & Engineering, Indian Institute of Technology (Banaras Hindu University), Varanasi. The matter embodied in this Ph. D. thesis has not been submitted for the award of any other degree/diploma.*

Date: 16th January 2018
Place: Varanasi **(Harish Kumar Shakya)**

Department of Computer Science& Engineering
Indian Institute of Technology
(Banaras Hindu University)
Varanasi – 221 005

ANNEXURE- E

CANDIDATE'S DECLARATION

*I, **Harish Kumar Shakya**, certify that the work embodied in this Ph.D. thesis is my own bonafide work carried out by me under the supervision of Dr Bhaskar Biswas for a period of about 06 years 06 Months from July, 2011 to January, 2018 at the Department of Computer Science & Engineering, Indian Institute of Technology (Banaras Hindu University), Varanasi. The matter embodied in this Ph.D. thesis has not been submitted for the award of any other degree/diploma.*

I declare that I have faithfully acknowledged, given credit to and referred to the research workers wherever their works have been cited in the text and the body of the thesis. I further certify that I have not willfully lifted up some other's work, para, text, data, results, etc. reported in the journals, books, magazines, reports, dissertations, theses, etc., or available at web-sites and included them in this Ph.D. thesis and cited as my own work.

Date: 16th January 2018
Place: Varanasi **(Harish Kumar Shakya)**

Certificate from the Supervisor/Co-supervisor/Head of the Department

This is to certify that the above statement made by the candidate is correct to the best of our knowledge.

(Dr. Bhaskar Biswas)
Supervisor

(Prof. Rajeev Srivastva)
Head of the Department

Department of Computer Science & Engineering
Indian Institute of Technology
(Banaras Hindu University)
Varanasi – 221 005

ANNEXURE- F

COURSE/COMPREHENSIVE EXAMINATION/PRE-SUBMISSION SEMINAR COMPLETION CERTIFICATE

This is to certify that ***Mr. Harish Kumar Shakya****, a bonafide research scholar of this Department, has satisfactorily completed the course work/comprehensive examination. The pre-submission seminar has successfully completed on 29.07.2017, entitled "Community Detection in Social Networks using Improvised Evolutionary Algorithms" which is a part of his Ph.D. programme.*

Date: 16th January 2018

Place: IIT BHU, Varanasi ***(Head of the Department)***

Department of Computer Science & Engineering
Indian Institute of Technology
(Banaras Hindu University)
Varanasi – 221 005

ANNEXURE- G

COPYRIGHT TRANSFER CERTIFICATE

Title of the Thesis: "Community Detection in Social Networks using Improvised Evolutionary Algorithms"

Candidate's Name: Mr. Harish Kumar Shakya

Copyright Transfer

Date: 16th January 2018 **(Harish Kumar Shakya)**

Acknowledgements

I would like to take the opportunity to thank everybody who inspired me, helped me and contributed directly or indirectly in realizing this thesis. Predominantly, my sincere thanks and regards goes to my thesis supervisor ***Dr. Bhaskar Biswas*** for his support and guidance during the execution of research. It would not have been possible to successfully carry out this work without his directions, advices and suggestions.

I would like to express my sincere thanks for ***Prof. Rajeev Srivastava***, Head, Computer Science and Engineering Department, for his kindness and valuable support in carrying out the research. I express my sincere thanks to the faculty members and staffs of the department, ***Prof. A. K. Tripathi, Prof. K. K. Shukla, Prof. R. B. Mishra, Prof. S. K. Singh, Dr. Vinayak Srivastava, Dr. Bhaskar Biswas***, ***Dr. Hari Prabhat Gupta*** and ***Dr Tanima Dutta for*** their extensive and inspiring guidance throughout this tenure.

I specially thank the members of my Research Progress Evaluation Committee, ***Prof. K. K. Shukla, Prof. Rajeev Srivastava, Prof. Neeraj Sharma,*** and DPGC Convener ***Dr. Bhaskar Biswas*** for their invaluable suggestions regarding the thesis.

I wish to express my deep sense of thanks to ***Mr. Kuldeep Singh***, ***Mr. Dharmendra Prasad Mahato***, ***Mr. Gyanendra Singh*** and ***Ms. Roli***, for being great friends and the best advisers I could ever have. Their advice, encouragement and critics were always the source of inspiration. This dissertation would not have been possible without their invaluable suggestions and persistent help. I wish to extend my gratitude towards all my lab mates ***Mr. Ajay Singh, Mr. Shashank Singh, Mr. Sumit Jaiswal,*** and ***Mr. Anupam Biswas*** whose company helped me a lot.

I would wish to thank all my family members who continuously supported me throughout my life. My parents did a lot of hard work to support the logistics for my studies and always encouraged me in every decision that I took. I would like to thank my Father, who was also my childhood teacher and is always there to stands by my side. I owe my thanks to my younger brother ***Mr. Ravi*** and sisters ***Ms. Hema & Mrs. Nisha*** who help me fulfill my obligation to take good care of our parents when I am far from home. I thank all of them for the sacrifice they made during this hard-working, yet fun-filled period in my life.

I would like to extend a warm thank to my wife ***Deepshikha*** and my loving son ***Nishchhal*** for all the sacrifice they have done for me in order to succeed throughout my life particularly during the completion of this.

I take this opportunity to sincerely acknowledge the University Grants Commission, Government of India for providing Ph.D. Fellowship for financial assistance. I sincerely thank all of them who contributed in for helping me to see the light at the end of every scary tunnel during my Ph.D.

Finally, I am grateful to God for his blessings and for giving me the strength to persevere throughout the long and arduous journey towards a Ph.D.

-Harish Kumar Shakya

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ACO	Ant Colony Optimization
ANS	Anticlockwise Negative Shift
ANUI	Average Normalized Unifiablity and Isolability
APS	Anticlockwise Positive Shift
ARI	Adjusted Rand Index
AUC	Area Under the receiver operating characteristic Curve
AVI	Average Isolability
AVU	Average Unifiability
CDF	Cumulative Density Function
CEW	Community-based Edge Weight
CFGC	Fuzzy Center-based Graph Clustering
CLP	Community-based Link Prediction
CNS	Clockwise Negative Shift
CPS	Clockwise Positive Shift
EA	Evolutionary Algorithm
EC	Edge Centrality
EOA	Evolutionary Optimization Algorithm
ERW-Kpath	Edge Random Walk k-Path
ExtD	External Density
FastU	Fast Unfolding
FMM/H2	Fuzzy Modularity Maximization/Heuristic 2
FN	False Negative
FP	False Positive
FGA	Fuzzy Genetic Algorithm
GA	Genetic Algorithm
GAFCD	Genetic Algorithm based Fuzzy Community Detection
GR-QC	General Relativity and Quantum Cosmology

GALS	Genetic Algorithm With Local Search
HC-PIN	Hierarchical Clustering-Protein Interaction Networks
HEP-TH	High Energy Physics - Theory
JDE	Janez Differential Evolution
LFR	Lancichinetti, Fortunato and Radicchi
MCOBGA	Modified Crossover and Opposition Based Genetic Algorithm
MGAFCD	Modified Genetic Algorithm for Fuzzy Community Detection
MSFCM	Multi-cut Spectral Fuzzy C-Means
MCDM	Multiple Criterion Decision Making
NL	Neutral Line
NMI	Normalized Mutual Information
NoC	Number of Communities
NSGAP	Node Similarity Based Genetic Algorithm with Permanence Concept
OBGA	Opposition Based Genetic Algorithm
OBDE	Opposition Based Differential Evolution
RandW	Random Walk
RL	Regression Line
RGA	Regenerative Genetic Algorithm
RS	Ranging Score
RW	Random Walk
SGA	Simple Genetic Algorithm
SCAN	Structural Clustering Algorithm for Networks
SIS	Susceptible-Infected-Susceptible
SLR	Simple Linear Regression
TN	True Negative
TOPSIS	Technique for Order Preference by Similarity to Ideal Solution
TP	True Positive
TDE	Tournament Based Differential Evolution
TOBDE	Tournament and Opposition Based Differential Evolution
VSDE	Vertex Similarity Based Differential Evolution
VGA	Vertex Similarity Based Genetic Algorithm

LIST OF SYMBOLS

C	detected community structure
R	real community structure
C_i	community i
α	threshold for Reachability
β	threshold for Isolability
η	ego network
μ	mixing parameter
ψ	membership threshold
U	partition matrix
U^+	extended partition matrix
t_{max}	maximum number of iterations
δ_i	size of community that include node i
$\sigma_{st}(ei\,j)$	number of shortest paths between node i and j
$\sigma^{k}_{st}(ei\,j)$	number of paths having at least k length
T_s	number of spanning trees
$\delta(u,v)$	strength of connection between any two nodes u and v

Preface

Complex networks such as social networks exhibit disproportionate connections among different nodes, resulting in densely interconnected groups of nodes. These highly connected groups of nodes within the network are referred as communities, which have significant role in understanding and uncovering various functional properties of the system. Today, introduction of social networking applications into every area of our lives makes social network analysis an important research area. An important property of networks/graphs modeling complex systems is the property of community structure, in which nodes are joined together in tightly knit groups (communities or clusters), between which there are only looser connections. The problem of detecting and extracting communities from such graphs has been the subject of intense investigations in recent years. This problem is very hard and not yet satisfactorily solved. Identification of communities has grown as one of the major research topics in social network analysis.

In this thesis, we explore and work on this community detection problem. We frame the problem as an optimization problem and hence explore the use of Evolutionary Algorithm i.e. Genetic Algorithms (GAs) & Differential Evolution (DE) in solving the same. We have studied, analyzed and implemented several existing algorithms including standard ones and GA-based ones. The standard algorithms include the Girvan-Newman (GN) Algorithm,FN, and the Label Propagation Algorithm by Raghavan et al. while the GA-based is Tasgin et al. algorithm ,Vertex similarity based MENSGA, GAFCD, GALS,TGA,CCGA,LGA, . We have also designed a new GA-based algorithm for the problem. For this, a modified Genetic Algorithm of which chromosome structure and genetic operators are modified to find communities in social networks is used. This modified Genetic Algorithm can be used without giving proposed community number at the initialization with OBL (opposition based learning) and it runs faster compared to other Genetic Algorithm methods. Additionally, we did some other modification in Genetic algorithm just like name as MCOBGA, RGA, FGA, MGAFCD, and NSGAP.

We present a comparative performance (accuracy + quality) analysis of these algorithms (new + existing) to gain insights into the problem and reveal the advantages of our proposed algorithm over existing algorithms. We have also created some artificial datasets (based on standard existing algorithms like the one for LFR graphs) for the purpose of the analysis and have acquired some real-world datasets (like Zachary's karate club network, American Football club, Strike dataset, Lusseau's network of bottlenose dolphins, etc.) too.

First two proposed chapter, we worked on the Genetic algorithm for the disjoint and overlapped community detection in social networks. We proposed a opposition based learning concept with genetic algorithm and called a MCOBGA, in this algorithm we find the disjoint communities for different datasets. Next one is the RGA; in this method we employed the regeneration of population behalf of mutation operation. In this algorithm, we update the convergence rate and efficiency and verified the number of artificial and real world datasets.

Another proposed chapter, we worked on the fuzzy community detection with the help of genetic algorithm. In FGA, we employed the fuzzy modularity concept (Liu & Zhang) and Omega for the overlapped community in social network. Similarly other one is modified the existing algorithm GAFCD; that is detect the crisp and fuzzy community both. So-that we modified this algorithm and create a new comparative method. In the last section of this proposed chapter, NSGAP means node similarity based genetic algorithm with permanence concept for the fuzzy and crisp community detection in social network.

In the last proposed chapter, we worked on the Differential evolution algorithm for the community detection in social network and find the best DE version of the different situation and the various types of datasets. We employed the tournament method; opposition based learning concept and vertex similarity concepts with DE algorithm. We created the TOBDE, OBDE, TDE, VSDE and the DE with Multiple objective functions.