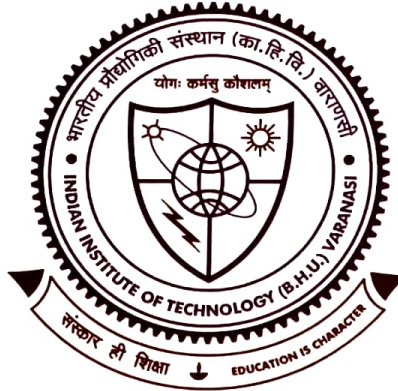


New Proposals for Community Detection in Dynamic Social Networks



Thesis submitted in partial fulfillment for the
Award of Degree
Doctor of Philosophy

By
Sneha Mishra

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY
(BANARAS HINDU UNIVERSITY)
VARANASI- 221005
INDIA

Roll No.: 17071009

June 2024

Certificate

It is certified that the work contained in this thesis entitled "New Proposals for Community Detection in Dynamic Social Networks" by "Sneha Mishra" has been carried out under my supervision and that it has not been submitted elsewhere for a degree.

Bhaskar
26/06/2024 -

Dr. Bhaskar Biswas

Associate Professor

Department of Computer Science and Engineering

Indian Institute of Technology (Banaras Hindu University),

Varanasi-221 005

सह आचार्य/Associate Professor
संगणक विज्ञान एवं अभियंत्रण विभाग/Department of Computer Sc. & Engg.
भारतीय प्रौद्योगिकी संस्थान/Indian Institute of Technology
बनारस हिन्दू विश्वविद्यालय/(Banaras Hindu University)

DECLARATION BY THE CANDIDATE

I, Sneha Mishra, certify that the work embodied in this thesis is my own bonafide work and carried out by me under the supervision of Dr. Bhaskar Biswas from July-2017 to June 2024, at the Department of Computer Science and Engineering, Indian Institute of Technology (BHU), Varanasi. The matter embodied in this thesis has not been submitted for the award of any other degree/diploma. I declare that I have faithfully acknowledged and given credits to the research workers wherever their works have been cited in my work in this thesis. I further declare that I have not willfully copied any other's work, paragraphs, text, data, results, etc., reported in journals, books, magazines, reports dissertations, theses, etc., or available at websites and have not included them in this thesis and have not cited as my own work.

Sneha Mishra

Date : 23/06/2024

Signature of the Student

Place : Varanasi

(Sneha Mishra)

CERTIFICATE BY THE SUPERVISOR

It is certified that the above statement made by the student is correct to the best of my/our knowledge.

Bhaskar 26/06/24

Signature of Supervisor

(Dr. Bhaskar Biswas)

पर्यवेक्षक/Supervisor
संगणक विज्ञान एवं अभियंताओं विभाग
Department of Computer Sc. & Engg
भारतीय प्रौद्योगिकी संस्थान
Indian Institute of Technology
(कानपुर हिन्दू विश्वविद्यालय)
(Banarus Hindu University)
वाराणसी-221005

Signature of Head of Department

आचार्य व विभागाध्यक्ष
Professor & Head
संगणक विज्ञान एवं अभियंताओं विभाग
Department of Computer Sc. & Engg
भारतीय प्रौद्योगिकी संस्थान
Indian Institute of Technology
(कानपुर हिन्दू विश्वविद्यालय)
वाराणसी-221005

COPYRIGHT TRANSFER CERTIFICATE

Title of the Thesis: New Proposals for Community Detection in Dynamic Social Networks

Name of the Student: Sneha Mishra

Copyright Transfer

The undersigned hereby assigns to the Indian Institute of Technology (Banaras Hindu University), Varanasi all rights under copyright that may exist in and for the above thesis submitted for the award of the Doctor of Philosophy.

Date: 25/06/2024

Sneha Mishra
(Sneha Mishra)

Place: Varanasi

Note: However, the author may reproduce or authorize others to reproduce material extracted verbatim from the thesis or derivative of the thesis for author's personal use provided that the source and the Institute's copyright notice are indicated.

Acknowledgements

In embarking on this intellectual journey, I find myself humbled by the myriad of individuals whose support, guidance, and encouragement have propelled me forward. To each of you, I extend my deepest gratitude and heartfelt appreciation.

First and foremost, I offer my sincerest thanks to the Almighty, whose grace and benevolence have been my guiding light, providing solace in moments of doubt and strength in times of adversity.

To my esteemed supervisor, Dr. Bhaskar Biswas, words cannot adequately express the depth of my gratitude for your unwavering mentorship, sage advice, and steadfast support. Your dedication to my academic and personal growth has been a beacon of inspiration, and I am truly privileged to have had the opportunity to learn from you.

I am equally indebted to the esteemed members of my RPEC committee, Dr. Ravi Sankar Singh, Computer Science and Engineering and Dr. M. Thottappan, Electronics Engineering, IIT(BHU), Varanasi, whose expertise, constructive criticism, and insightful feedback have enriched the quality of this thesis immeasurably. Your commitment to excellence has challenged me to strive for the highest standards, and for that, I am profoundly grateful.

To the Head of the Department, Prof. Sanjay Kumar Singh, I extend my heartfelt appreciation for your visionary leadership and unwavering support of academic pursuits. Your encouragement and advocacy have created an environment conducive to intellectual exploration and scholarly achievement.

A heartfelt thank you is also extended to the dedicated office staff whose administrative assistance and logistical support have facilitated the smooth execution of this endeavor. Your professionalism and efficiency have not gone unnoticed, and I am deeply appreciative of your contributions.

To my esteemed lab mates, Dr. Shanshank S. Singh, Dr. Shivansh Mishra, Dr. Vishal Srivastav, Dr. Ajay Kumar and Mr. Naveen M. Upadhyay, your camaraderie, intellectual exchange, and collaborative spirit have enriched my research experience beyond measure. The shared triumphs and challenges we have encountered together have been a source of inspiration and camaraderie that I will cherish always.

I would be remiss not to acknowledge the unwavering support of my dear friends, whose encouragement, understanding, and unwavering belief in my abilities have sustained me through the most challenging of times. Your friendship is a cherished gift, and I am immensely grateful for your presence in my life.

To my beloved father, Dr. Ramesh Chandra Mishra, mother, Mrs. Nirmala Mishra and spouse Dr. Shwet Ketu, your boundless love, sacrifices, and unwavering belief in my potential have been the cornerstone of my success. Your enduring support and encouragement have been a source of strength and inspiration, and I am forever grateful for the values you have instilled in me.

I would like to put special thank to Swati Mishra and Anushka for cheering me in all my low. Lastly, to my cherished family members, whose love, encouragement, and unwavering belief in my abilities have been a constant source of motivation and inspiration, I extend my heartfelt gratitude. Your unwavering support has been the wind beneath my wings, propelling me forward on this remarkable journey.

In concluding this acknowledgment, I am reminded of the words of Helen Keller, who once said, "Alone we can do so little; together we can do so much." Indeed, it is with profound humility and gratitude that I acknowledge the collective contributions of each individual mentioned herein, for it is through your collective support and encouragement that this achievement has been made possible.

Sneha Mishra

- Sneha Mishra

Contents

Certificate	ii
Acknowledgements	v
Contents	vii
List of Figures	xi
List of Tables	xiii
Abbreviations	xv
Symbols	xvii
Abstract	xix
1 Introduction	1
1.1 Background	1
1.2 Open Issues	4
1.3 Challenges in Community Detection in Dynamic Networks	4
1.4 Applications of Community Detection	6
1.5 Objectives	8
1.6 Contribution	10
1.6.1 Tree based Community Detection	10
1.6.2 Multi-objective based Community Detection	10
1.6.3 Membership based Community Detection	11
1.7 Thesis Organization	11
2 Related Work	13
2.1 Introduction	13

2.2	Dynamic Network	14
2.2.1	Network Model	15
2.2.1.1	Temporal Dynamic Network	15
2.2.1.2	Snapshot Dynamic Network	16
2.3	Related Literature	17
2.3.1	Graph Representation Based	18
2.3.1.1	Matrix Factorization Based	19
2.3.1.2	Neural Network Based	20
2.3.2	Optimization Based	20
2.3.2.1	Single-objective	22
2.3.2.2	Multi-objective	22
2.3.3	Technique Based	24
2.3.3.1	Clique Percolation Method	24
2.3.3.2	Non-negative Matrix Factorization	26
2.3.3.3	Random Walk	27
2.3.3.4	Tree Based	28
2.3.3.5	Label propagation	28
2.4	Evaluation Metrics	29
2.4.1	Precision and Recall	29
2.4.2	F Score	31
2.4.3	Normalized Mutual Information	31
2.4.4	Adjusted Rand Index	32
2.4.5	Entropy	33
2.4.6	Purity	33
2.4.7	Average Isolability	33
2.4.8	External Density	34
2.4.9	Coverage	34
2.4.10	Modularity	34
3	Tree based Community Detection	35
3.1	Introduction	35
3.2	Proposed Work	37
3.2.1	Algorithm	39
3.2.2	Applying the Algorithm	41
3.2.3	Complexity Analysis	42
3.3	Evaluation Setup	43
3.3.1	Performance Metrics	43
3.3.2	Datasets	43
3.3.2.1	Known Community Structures	44
	Real World Datasets	44
	Synthetic Datasets	44

3.3.2.2	Unknown Community Structures	47
3.3.3	Methods for Comparison	56
3.4	Performance Analysis	57
3.4.1	Parameter Analysis	57
3.4.1.1	Window size W and dissolution threshold θ	57
3.4.1.2	Mixing parameter μ	57
3.4.2	Accuracy Analysis	58
3.4.3	Quality Analysis	58
3.4.4	Performance Measuring under Dynamic Settings	59
3.5	Conclusion and Future Work	59
4	Multiobjective based Community Detection	61
4.1	Introduction	61
4.2	Community Life Cycle	64
4.3	Multi-objective Optimization	66
4.4	Proposed Work	67
4.4.1	Mathematical Formulation	68
4.4.2	Dynamic Settings	70
4.4.3	Algorithm	71
Candidate identification	73	
Solution space formation	73	
Pareto solution	73	
4.4.4	Complexity Analysis	74
4.5	Performance Assessment	75
4.5.1	State-of-the-art Algorithms	75
4.5.2	Datasets	77
4.5.3	Metrics	77
4.6	Results	78
4.6.1	Quality	79
4.6.2	Accuracy	83
4.7	Discussion	86
4.8	Conclusion and Future Work	87
5	Membership based Disjoint Community Detection	89
5.1	Introduction	89
Contribution	90	
5.2	Proposed Work	91
5.2.1	Membership Function	91
5.2.2	Algorithm	92
5.2.3	Complexity Analysis	95
5.3	Performance Assessment	96
5.3.1	State-Of-The-Art Algorithms	96

Contents

5.3.2	Datasets	97
5.3.2.1	Real Datasets	97
5.3.2.2	Synthetic datasets	98
5.3.3	Metrics	98
5.3.4	Results	99
5.3.4.1	Quality	102
5.3.4.2	Accuracy	105
5.4	Conclusion and Future Work	105
6	Conclusion and Future Research Direction	107
6.1	Conclusion	107
6.1.1	Key Research Findings	108
6.2	Future Directions	110
A	List of Publications	113
	Bibliography	115

List of Figures

1.1	A schematic representation of a network with disjoint and overlapping community structure	2
1.2	Explanation of thesis organization in relation to the four main contributory chapters	11
2.1	Dynamic network models in increasing order of temporal granularity . . .	15
2.2	Temporal network model for a dynamic network	17
2.3	Snapshot network model for a dynamic network	17
2.4	Dynamic network models in increasing order of temporal granularity . . .	18
3.1	The proposed framework TCD2	37
3.2	The identification of communities on example graph based on expansion phase of TCD2	43
3.3	Safe zone predicted with different metric values corresponding to w and θ ranges over 0.2 – 1.0 and 0.1 – 0.5 respectively based on three-degree theory (level=3) in strike dataset	46
3.4	The comparison of different metric values corresponding to mixing parameter μ based on three degree theory (level=3) and six degree of separation (level=6) in synthetic dataset	47
3.5	Safe zone predicted with different metric values corresponding to w and θ ranges over 0.2 – 1.0 and 0.1 – 0.5 respectively based on three degree theory (level=3) in karate dataset	48
3.6	Safe zone predicted with different metric values corresponding to w and θ ranges over 0.2 – 1.0 and 0.1 – 0.5 respectively based on three degree theory (level=3) in dolphin dataset	49
3.7	Safe zone predicted with different metric values corresponding to w and θ ranges over 0.2 – 1.0 and 0.1 – 0.5 respectively based on three degree theory (level=3) in football dataset	50
3.8	Safe zone predicted with different metric values corresponding to w and θ ranges over 0.2 – 1.0 and 0.1 – 0.5 respectively based on small world phenomena (level=6) in Gr-QC dataset whose ground truth communities are unknown	51

3.9	Safe zone predicted with different metric values corresponding to w and θ ranges over 0.2 – 1.0 and 0.1 – 0.5 respectively based on small world phenomena (level=6) in HEP-TH dataset whose ground truth communities are unknown	51
3.10	Safe zone predicted with different metric values corresponding to w and θ ranges over 0.2 – 1.0 and 0.1 – 0.5 respectively based on small world phenomena (level=6) in Wiki-Vote dataset whose ground truth communities are unknown	52
4.1	Community life cycle	65
4.2	The working of proposed algorithm	74
4.3	The evolution of community structure in an example graph at different time stamps	75
5.1	Flow chart of algorithm	93

List of Tables

2.1	Event table for temporal network model	16
2.2	Relevant literature for graph representation based community detection in dynamic network	21
2.3	Relevant literature for objective based community detection in dynamic network	25
2.4	Relevant literature for technique based community detection in dynamic network	30
2.5	Contingency table	32
3.1	List of rules during INSERTION operations in expansion phase of TCD2 framework, when a edge (x,y) is identified in dynamic setting (0: indicates new node is identified in the network; 1: indicates node is already presents on the network)	38
3.2	The updation of tree-structure using Expansion Phase on exemplary graph	42
3.3	Safe range of window size and θ for quality and accuracy metrics corresponding to datasets	45
3.4	The Comparison of Accuracy Metric Values of TCD2 Algorithm against the state-of-the-art algorithms in various datasets whose ground truth communities known	52
3.5	The Comparison of Quality Metric Values of TCD2 Algorithm against the State-of-the-art Algorithms in Various Datasets	53
3.6	The Comparison of Accuracy Metric Values of TCD2 Algorithm under Dynamic Settings (at Different Timestamps)	54
3.7	The Comparison of Quality Metric Values of TCD2 Algorithm under Dynamic Settings (at Different Timestamps)	55
4.1	Various stages of community life cycle in evolution of community structure in the exemplary graph	76
4.2	Datasets used in performance evaluation	78
4.3	Comparison of quality metrics value of state-of-the-art methods on connected datasets	80
4.4	Comparison of quality metrics value of state-of-the-art methods on disconnected datasets	82

4.5	Comparison of accuracy metrics value of state-of-the-art methods on connected datasets	84
4.6	Comparison of accuracy metrics value of state-of-the-art methods on disconnected datasets	85
5.1	Quality metrics results' comparison of the algorithm with state of the art algorithms	99
5.2	Accuracy metrics results' comparison of the algorithm with state of the art algorithms	103

Abbreviations

Abbreviation	Description
CD2N	Community Detection in Dynamic Network
TCD2	Tree Based Communnity Detection in Dynamic social network
MDCD	Membership based Disjoint Community Detection
DN	Dynamic Network
TNM	Temporal Dynamic Network
SNM	Snapshot Network Model
GNN	Graph Neural Network
GCN	Graph Convolutional Network
RNN	Recurrent Neural Network
LSTM	Long Short Term Memory
GRU	Gated Recurrent Unit
GAT	Graph Attention Network
GAN	Generative Adversarial Network
CNN	Convolutional Neural Network
PSO	Particle Swarm Optimization
MOP	Multi-objective Optimization
SOO	Single Objective Optimization
CPM	Clique Percolation Method
NMF	Non-negative Matrix Factorization
LPA	Label Propagation Algorithm

NMI	Normalized Mutual Information
ARI	Adjusted Rand Index
LFR	Lancichinetti-Fortunato-Radicchi benchmark
CLC	Community Life Cycle

Symbols

$G_t(V_t, E_t)$	Graph at time t
V_t	Set of nodes at time t
E_t	Set of edges at time t
\mathbf{C}_t	Set of communities at time t
δ	Time duration of an event
Δ	Time duration between two snapshots
θ	Dissolution threshold
l	Number of levels for tree structure
W	Window size for dissolution phase
μ	Mixing parameter
C_t^i	t^{th} community C at time t
C_t^{can}	Candidate community set
$N(u)$	Neighbour set of node u
$CN(u, v)$	Common neighbours of u and v
w_{vx}	Weight of edge (v, x)

ABSTRACT

A network is fundamentally composed of nodes, which represent the actors within it, and edges, which represent the relationships between these actors. A community in a network is a collection of entities that share a significant number of closely related interactions. Detecting these communities is a critical area of research within the field of data mining, as evidenced by extensive literature on the subject. Social networks that evolve over time present unique challenges; they demand algorithms that can accommodate temporal changes while operating within limited resource constraints. Static algorithms typically fall short in such dynamic environments, where there is a need for continuous updates to community structures, efficient computations, and immediate results.

Dynamic networks introduce new challenges because they are characterized by their ability to evolve, necessitating models that account for both their structural and temporal patterns. Research indicates that these networks, which reflect real-world scenarios, are organized into communities that frequently change over time. These communities can represent tightly-knit groups such as linked online documents, social circles, or even consumer purchasing patterns.

This thesis presents three novel techniques for the identification of communities within dynamic social networks. Prior to delving into the algorithms, an exposition on a network model tailored for dynamic networks is presented. The network is conceptualized as an evolving graph, depicted as a stream of events. Each event signifies a modification within the network, accompanied by corresponding timestamps.

The first method utilizes a tree-based algorithm, TCD2, which takes advantage of social network properties like connectedness and influence to detect communities. Nodes are categorized in two roles: cluster heads and members. Further they are organized in the tree where the root node stores cluster heads and their corresponding members are stored at

subsequent levels. TCD2 employs a tree structure to keep track of the network's changing community structures.

Subsequently, the thesis details a multi-objective optimization approach for community detection in dynamic networks, viewing each network as a temporal sequence of events characterized by the introduction of new edges. This method updates the community affiliations of nodes based on the recency of interactions, utilizing the Pareto front principle to determine the most fitting community.

Finally, a fuzzy logic approach to community detection in dynamic networks is presented, featuring a novel membership function based on insights from the literature. This function is critical in assigning nodes to communities at the decision-making step, enabling the algorithm to generate distinct communities at any given point in time.

Each of the aforementioned techniques has been thoroughly evaluated against the recognized metrics from the literature. Both quality and accuracy metrics are used in the evaluation of proposed algorithms. The experimental results on real world social networks along with synthetic networks validate the performance of algorithms. An empirical analysis has also been included that compares the effectiveness of the proposed algorithms with the current state-of-the-art algorithms, demonstrating their potential and performance. Each algorithm has the capability to delineate the community structure of a given dynamic network at a specific point in time.

This thesis is dedicated to my beloved family...