Link Prediction in Dynamic Networks using Feature and Quantum based Machine Learning Techniques

Thesis submitted in partial fulfillment

for the Award of Degree

DOCTOR OF PHILOSOPHY

By

MUKESH KUMAR

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY

(BANARAS HINDU UNIVERSITY),

VARANASI-221 005

Roll No: 19071009 2023

Certificate

It is certified that the work contained in this thesis entitled "Link Prediction in Dynamic Networks using Feature and Quantum based Machine Learning Techniques" by "MUKESH KUMAR" has been carried out under my supervision and that this work has not been submitted elsewhere for a degree.

It is further certified that the student has fulfilled all the requirements of Comprehensive Examination, Candidacy and SOTA for the award of Ph.D. Degree.

Bhaskar 30/01/2023

Dr. Bhaskar Biswas

Associate Professor

Department of Computer Science and Engineering

Indian Institute of Technology (Banaras Hindu University),

Varanasi-221 005

COPYRIGHT TRANSFER CERTIFICATE

Title of the Thesis : **Link Prediction in Dynamic Networks using Feature and Quantum based Machine Learning Techniques**

Name of the Student : **MUKESH KUMAR**

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 : 30/01/2023.

Place : Varanasi

Mukesh Kumar

Signature of the Student

(MUKESH KUMAR)

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.

DECLARATION BY THE CANDIDATE

I, MUKESH KUMAR, 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 29, 2019 to January 29, 2023, 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.

Date :

Place : Varanasi

Signature of the Student

(MUKESH KUMAR)

CERTIFICATE BY THE SUPERVISOR

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

30/01/2023

Signature of Supervisor

(Dr. Bhaskar Biswas)

पर्यवेक्षक/Supervisor
संगणक विज्ञान एवं अभियांत्रिकी विभाग
Department of Computer Sc. & Engg
भारतीय प्रौद्योगिकी संस्थान
Indian Institute of Techno
(काशी हिन्दू विश्वविद्या
(Banaras Hindu Universi
Varanasi-22100

Signature of Head of Department

आचार्य व विभागाध्यक्ष
Professor & Head
संगणक विज्ञान एवं अभियांत्रिकी विभाग
Department of Computer Sc. & Engg
भारतीय प्रौद्योगिकी संस्थान
Indian Institute of Technology

Acknowledgements

This thesis is the end of my journey in obtaining my Ph.D. I have not traveled in a vacuum in this journey. This thesis has been kept on track and been seen through to completion with the support and encouragement of numerous people including my well wishers, my friends, colleagues and various institutions. At the end of my thesis, I would like to thank all those people who made this thesis possible and an unforgettable experience for me. At the end of my thesis, it is a pleasant task to express my thanks to all those who contributed in many ways to the success of this study and made it an unforgettable experience for me.

Firstly, I would like to express my sincere gratitude to my advisor **Dr. Bhaskar Biswas** for the continuous support of my Ph.D study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D study.

I specially thank the members of my Research Progress Evaluation Committee, **Dr. Sanjay kumar Singh** and **Dr. Rajeev** for their invaluable suggestions regarding the thesis, their insightful comments and encouragement, but also for the hard question which incented me to widen my research from various perspectives.

I would like to express my sincere thanks for **Prof. Sanjay Kumar Singh**, 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, **Prof. K. K. Shukla**, **Prof. A. K. Tripathi**, **Dr. Ravi Shankar Singh**, **Dr. Vinayak Srivastava**, **Dr. R. N. Choudary**, **Dr. H. P. Gupta**, **Dr. T. Dutta**, **Dr. A. K. Singh** and staffs of the department.

I am also thankful to all my friends for the fun moments in my PhD student life. I wish to express my deep sense of thanks to my senior **Dr. Ajay Kumar** and **Mr. Shivansh Mishra** for being a 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.

Most importantly, my deepest gratitude is for my family for their constant support, inspiration, guidance, and sacrifices. My parents were constant source of motivation and inspiration. Their affection and guidance was instrumental in me choosing Engineering and eventually continuing on to my Ph.D. I would like to thank my mother, who was also my childhood teacher and is always there to stands by my side.

Last but not the least, I would like to special thanks is to my wife **Miss Anjali Kumari**, my daughters **Srishti Yadav and Trisha** and my son **Shivansh Yadav** for supporting me unconditionally throughout my study. 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.

- MUKESH KUMAR

PREFACE

Link prediction is a task that involves network structure evaluation and identifying missing and future links in social networks. In dynamic networks, each edge is associated with a temporal component, and link prediction in such networks is the task of identifying edge existence probability based on previously observed network dynamics. This research is crucial to comprehending network evolution and its effects on individual node behavior with respect to time. Link prediction algorithms can be applied to extract missing data, identify erroneous interactions, evaluate the factors governing network evolution, and more. Node pair similarity-based link prediction methods use different structural or topological information to calculate possible edges' chance. In the snapshot-based paradigm for link prediction on dynamic graphs, feature sets created from combining different similarity-based methods are used to track the temporal changes in edges. These features are then used to make predictions for non-existing edges. Machine learning algorithms are used on these feature sets to build prediction models. Compared to the individual similarity-based approaches, machine learning techniques have been used to improve prediction accuracy as they are more helpful in modeling changes in patterns over time. This thesis aims to enhance link prediction on dynamic networks using different feature sets and machine learning algorithms.

The following four fundamental aspects of link prediction are explored in this thesis since they are used to build the feature sets for machine learning classifiers:

1) incorporating local, global, and quasi-local similarity indices for link prediction, 2) merging snapshot-based features with snapshot-impartial path-based and cost-based features, 3) feature selection to identify best feature sets based on similarity indices, 4) exploring community detection enhanced link prediction features, and 5) studying the effect of quantum kernel-inspired transformation on link prediction in dynamic networks.

One of the well-known categories of link prediction methods is the similarity-based method, which uses a node pair similarity score for edge estimation. The three widely used categories of similarity-based indices are Local (L), Global (G), and Quasi-local (Q) indices. All three types of these methods can be combined to create a rich feature set that would be more useful for link prediction than its components individually. The proposed LGQ framework explores the relative effect of combining these categories of link prediction features on different machine learning algorithms.

Further exploring the role of snapshot-impartial paths, Path Weight Aggregation Feature (PWAF) for Link Prediction in Dynamic Networks is proposed. Different topological aspects of the networks (Local, Global, and Quasi-local), as well as Clustering Coefficient based features, are considered for feature generation, in addition to the suggested Path Weight-Based Aggregation Feature (PWAF). Some of these features also take advantage of longer paths between nodes for a more thorough similarity estimation.

A feature-based solution that considers individual snapshots and the overall network throughout the full-time span to answer the link prediction problem is explored. A novel feature called Cost-based feature for link prediction (CFLP) for estimating edge behavior throughout the entire network is proposed, which uses a reward and penalty structure to summarize node activity across the whole network (snapshot-impartial), is presented. The feature set also uses similarity indices, classified into four major categories: local similarity, global similarity, quasi-local similarity, and clustering coefficient-based similarity, to measure edge activity change in individual snapshots. Different feature selection methods are used to correctly quantify the relative effect of features among themselves and the overall link prediction problem.

The highly connected groups of nodes within the network are called communities, which have an important role in understanding and uncovering various functional properties of the system. A community information-based feature estimation and link prediction (COMMLP) method applied to dynamic graphs in a per snapshot feature estimation-based setting is proposed to take advantage of the group behavior of nodes. First, a link prediction framework is presented to predict missing links using parameterized influence regions of nodes and their contribution to community partitions. Then, a unique feature set is generated using local, global, and quasi-local similarity-based and community information-based features. This feature set is further optimized using scoring-based feature selection methods to select only the most relevant features.

Finally, the solution to the issue of link prediction in dynamic networks using supervised learning and the Projected Quantum Kernel (PQK) is addressed. It is conceivable to expect that quantum computers may outperform classical computers in machine learning tasks since quantum systems display aberrant behavior that conventional systems are thought to be incapable of creating. The QML approaches use the huge dimensionality of quantum Hilbert space to get an optimized solution by modeling the feature space of a classification problem with a quantum state. The projected quantum kernel-based link prediction (PQKLP) approach is presented to address the link prediction problem using both local and global information. Using PQK, features are transformed from the popularly used snapshot-based feature set form into quantum space such that the effectiveness of machine learning-based classification can be improved.

Contents

List of Figures

List of Tables

Abbreviations

L	**L**ocal **S**imilarity **I**ndices
G	**G**lobal **S**imilarity **I**ndices
Q	**Q**uasi-local **S**imilarity **I**ndices
LG	**L**ocal and **G**lobal **S**imilarity **I**ndices
LQ	**L**ocal and **Q**uasi-Local **S**imilarity **I**ndices
GQ	**G**lobal and **Q**uasi-Local **S**imilarity **I**ndices
LGQ	**L**ocal, **G**lobal and **Q**uasi-**L**ocal **S**imilarity **I**ndices
NN	**N**eural **N**etwork
LR	**L**ogistic **R**egression
XGB	**XGB**oost
LDA	**L**inear **D**iscriminant **A**nalysis
RFC	**R**andom **F**orest **C**lassifier
GNB	**G**aussian **N**aive **B**ays
CCLP2	**Level-2** Node **C**lustering **C**oefficient-based **L**ink **P**rediction
AUROC	**A**rea **U**nder the **R**eceiver **O**perating **C**haracteristic Curve
AP	**A**verage **P**reccision
AUPR	**A**rea **U**nder the **P**recision-**R**ecall Curve
L3	Path of length 3
CCLP	**C**lustering **C**oefficient-based **L**ink **P**rediction
PWAF	**P**ath **W**eight **A**ggregation **F**eature
CFLP	**C**ost **B**ased **F**eature for **L**ink **P**rediction
PQKLP	**P**rojected **Q**uantum **K**ernel based **L**ink **P**rediction

Symbols

$G(V,E)$	A social network with vertex set V and edge set E
A	Adjacency matrix of a network
n	The number of nodes in the network ($\|V\|$)
m	The number of Edges in the network ($\|E\|$)
$\Gamma(z)$	The neighbors set of node z
k_z	Degree of the node z
$t(z)$	Number of triangle passing through the node z
λ_1	Maximum eigen value of a matrix
β	dumping factor
$\langle K \rangle$	Average degree of the a network
$\langle D \rangle$	Average path length of a network
$\langle C \rangle$	Average clustering coefficient of a network
r	Coefficient of assortativity of a network
H	Degree of heterogeneity of a network
ρ	Network density
D	Diagonal matrix
L	Laplacian matrix
$C(z)$	Clustering coefficient of node z
CC(z)	Level-2 clustering coefficient a node z
$S(x,y)$	Similarity score between the node x and the node y