

Certificate

It is certified that the work contained in this thesis entitled “Enhanced Link Prediction using Aggregation for Edge Relevance Quantification” by “SHIVANSH MISHRA” 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.

Bhadkar
9/05/23

Dr. Bhaskar Biswas

पर्यवेक्षक/Supervisor
संगणक विज्ञान एवं अभियांत्रिकी विभाग
Associate Professor
भारतीय प्रौद्योगिकी संस्थान
Department of Computer Science and Engineering
Indian Institute of Technology (Banaras Hindu University),
Varanasi-221 005

Varanasi-221 005

DECLARATION BY THE CANDIDATE

I, SHIVANSH 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 26, 2016 to April 30, 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 : 01-05-2023

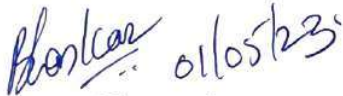
Signature of the Student

Place : Varanasi

(SHIVANSH 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.



Signature of Supervisor

पर्यवेक्षक/Supervisor
(Dr. Bhaskar Biswas)
Department of Computer Sc. & Engg
भारतीय प्रौद्योगिकी संस्थान
Indian Institute of Technology
(काशी हिन्दू विश्वविद्यालय)
(Banaras Hindu University)
वाराणसी/Varanasi-221005



Signature of Head of Department

आचार्य व विभागाध्यक्ष
Professor & Head
समणक विज्ञान एवं अभियान्त्रिकी विभाग
Department of Computer Sc. & Engg
भारतीय प्रौद्योगिकी संस्थान
Indian Institute of Technology
(वाराणसी हिन्दू विश्वविद्यालय)
(Banaras Hindu University)
वाराणसी-221005 / Varanasi-221005

COPYRIGHT TRANSFER CERTIFICATE

Title of the Thesis : **Enhanced Link Prediction using Aggregation for Edge Relevance Quantification**

Name of the Student : **SHIVANSH 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 : 01-05-2023

Place : Varanasi



Signature of the Student

(SHIVANSH MISHRA)

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

This thesis has been kept on track and been seen through to completion with the support and encouragement of numerous people, including well-wishers, friends, colleagues, and various institutions. With immense pleasure, I express my gratitude to all those who contributed in numerous ways towards the success of this study.

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

I am thankful to the members of my Research Progress Evaluation Committee, **Dr. R. S. Singh** and **Prof. Subir Das** for their invaluable suggestions regarding the thesis, their insightful comments and encouragement, but also for their critical viewpoints which required me to widen my research from various perspectives.

I would like to express my sincere thanks for **Prof. S. K. 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**, **Prof. Rajeev Srivastava**, **Dr. H. P. Gupta**, **Dr. A. Chaturvedi**, **Dr. Pratik Chattopadhyay**, **Dr. Ajay Pratap**, and supporting staff of the department. The valuable suggestions and constant support of **Prof. P. Chakrabarti**, **Dr. Karm Veer Singh** and **Dr. Pravin Singh Rana**, has also been instrumental towards the completion of this thesis.

I am also thankful to all my friends for their unwavering support during my Ph.D. I wish to express my deep gratitude to my seniors **Dr. Shashank Sheshar Singh**, **Mr. Mukesh Kumar**, **Dr. Ajay Kumar**, **Dr. Vishal Srivastava**, **Dr. Kuldeep Singh**, **Mr. S. P. Dwivedi** and **Mr. Naveen Upadhyay**, for being great friends, collaborators, and the best advisers I could ever have. Their advice, encouragement, and critique during the duration of this dissertation are much appreciated. This dissertation was only possible with their invaluable suggestions and persistent help. I am thankful to my childhood friends, **Mr. Saurabh Pandey**, **Mr. Akshay Pandey**, **Mr. Prajay Shukla**, **Mr. John Abdi**, and **Mr.**

Rohit Upadhyay, who have stuck together with me through thick and thin. I am also thankful for the discussions with my batchmates **Mr. Vinod Kumar** and **Mr. Supriya Chanda**, which has helped broaden my horizons in other research areas. I am especially thankful to **Dr. Shashank Sheshar Singh**, without whom this study might not even be possible. **Mr. Mukesh Kumar** has been my close collaborator and pillar of support throughout the tough times of the COVID-19 pandemic.

Most importantly, my deepest gratitude is to my family for their constant support, inspiration, guidance, and sacrifices. My parents, **Dr. Shraddha Upadhyaya** and **Dr. Jai Krishna Mishra** were a constant source of motivation and inspiration. Their affection and guidance were instrumental in getting me through the demanding duration of this thesis.

Last but not the least, I would like to praise and thank *Baba Kashi Vishwanath Mahadev*, the one who blessed me with the ability to undertake and finally complete this work.

- SHIVANSH MISHRA

List of Figures

| | | |
|-----|---|----|
| 1.1 | The Link Prediction (LP) finds missing links (i.e., AD, AC, and BD) in this observed network. | 4 |
| 1.2 | Link Prediction in Multiplex network (dashed line in Predictions for Layer-1) | 4 |
| 3.1 | Ego network structure | 40 |
| 3.2 | Ego network structure of node A (Lev.1 is Level-1(A) and so on and so forth) demonstrating the influence node A exerts on edges 3 hop distances away from it. | 42 |
| 3.3 | Example of different regions of feature selection for nodes X&Y (orange - CAR, red - CN, blue - CC, green - PA, black - Ego-2.0). | 45 |
| 3.4 | Example Network for demonstrating the working of <i>ELP</i> algorithm for link prediction. | 48 |
| 3.5 | Ego Strength of Example Network. | 49 |
| 3.6 | Accuracy Score comparison of <i>ELP</i> variations for different feature sets on six datasets. | 54 |
| 3.7 | AUPR comparison of <i>ELP</i> variations for different feature sets on six datasets. | 55 |
| 3.8 | AUC comparison of <i>ELP</i> variations for different feature sets on six datasets. | 57 |

| | | |
|-----|--|-----|
| 4.1 | Taxonomy of Path-based approaches to Link Prediction (Ajay et al.[1]) | 67 |
| 4.2 | HOPLP-MUL Concept Graph showing regions of influence of nodes X & Y with intersections between them which denote possible information flow paths. | 68 |
| 4.3 | Example Graph for demonstrating working of <i>HOPLP – MUL</i> for likelihood calculation over higher order paths. | 77 |
| 4.4 | AUC comparison of <i>HOPLP – MUL</i> variations for different feature sets on six datasets | 81 |
| 4.5 | F1 Score comparison of <i>HOPLP – MUL</i> variations for different feature sets on six datasets | 83 |
| 4.6 | Balanced Accuracy Score comparison of <i>HOPLP – MUL</i> variations for different feature sets on six datasets | 84 |
| 4.7 | Variation of algorithm performance of <i>HOPLP – MUL</i> with respect to γ on six datasets | 86 |
| 5.1 | Edge level network structure of node A (Lev.1 is Level-1(A) and so on and so forth) demonstrating the influence node A exerts on edges 3 hop distances away from it. | 103 |
| 5.2 | <i>MNERLP – MUL</i> Framework demonstrating the overall structure of link prediction workflow. | 105 |
| 5.3 | Example graph of <i>MNERLP – MUL</i> based link prediction for calculation of $LI(X, Y)$ | 110 |
| 5.4 | Heatmap of AUC variation of <i>MNERLP – MUL</i> algorithm's performance with respect to α & β with <i>Ratio</i> of testing to total edges averaged over range 0.1 – 0.5 | 112 |
| 5.5 | Heatmap of Balanced Accuracy score variation of <i>MNERLP – MUL</i> algorithm's performance with respect to α & β with <i>Ratio</i> of testing to total edges averaged over range 0.1 – 0.5 | 112 |

| | | |
|-----|---|-----|
| 5.6 | Heatmap of F1 score variation of <i>MNERLP – MUL</i> algorithm’s performance with respect to α & β with <i>Ratio</i> of testing to total edges averaged over range 0.1 – 0.5 | 113 |
| 5.7 | Graphs of AUC variation of <i>MNERLP – MUL</i> algorithm’s performance with respect to different Node Relevance (Centrality) variations | 114 |
| 5.8 | Graphs of Balanced Accuracy score variation of <i>MNERLP – MUL</i> algorithm’s performance with respect to different Node Relevance (Centrality) variations | 115 |
| 5.9 | Graphs of F1 score variation of <i>MNERLP – MUL</i> algorithm’s performance with respect to different Node Relevance (Centrality) variations | 116 |
| 6.1 | <i>CLP – MUL</i> Framework | 133 |
| 6.2 | Example Graph for <i>CLP – MUL</i> framework | 140 |
| 6.3 | AUC comparison of <i>CLP – MUL</i> algorithm for different feature sets on datasets | 143 |
| 6.4 | F1 Score comparison of <i>CLP – MUL</i> algorithm for different feature sets on datasets | 144 |
| 6.5 | Balanced Accuracy Score comparison of <i>CLP – MUL</i> algorithm for different feature sets on datasets | 146 |

List of Tables

| | | |
|-----|--|----|
| 2.1 | Statistical information of real-world datasets | 25 |
| 2.2 | Datasets and their basic topological characteristics | 31 |
| 3.1 | Feature Set Selection | 50 |
| 3.2 | Likelihood Score Computation | 51 |
| 3.3 | Running Time Analysis (in seconds) for different <i>Ratio</i> values representing testing to total edges percentage in five datasets. | 52 |
| 3.4 | Comparison of the proposed algorithm <i>ELP</i> with the state-of-the-art algorithms in terms of Accuracy Score | 59 |
| 3.5 | Comparison of the proposed algorithm <i>ELP</i> with the state-of-the-art algorithms in terms of AUPR | 60 |
| 3.6 | Comparison of the proposed algorithm <i>ELP</i> with the state-of-the-art algorithms in terms of AUC | 61 |
| 3.7 | The Posthoc Friedman Siegel Test (Control method = <i>ELP</i>) corresponding different metrics | 62 |
| 4.1 | SCORE and PRIOR matrices from Example for path-length $l = 2, 3, 4$. . . | 78 |

| | | |
|------|--|----|
| 4.2 | Comparison of the proposed algorithm <i>HOPLP – MUL</i> with baseline algorithms in terms of AUC on six datasets and five <i>Ratio</i> values for testing to total edges percentage | 88 |
| 4.3 | Comparison of the proposed algorithm <i>HOPLP – MUL</i> with baseline algorithms in terms of F1 Score on six datasets and five <i>Ratio</i> values for testing to total edges percentage | 89 |
| 4.4 | Comparison of the proposed algorithm <i>HOPLP – MUL</i> with baseline algorithms in terms of Balanced Accuracy Score on six datasets and five <i>Ratio</i> values for testing to total edges percentage | 90 |
| 4.5 | Comparison of the proposed algorithm <i>HOPLP – MUL</i> with baseline algorithms in terms of AUC layer-wise on three datasets and five <i>Ratio</i> values for testing to total edges percentage | 92 |
| 4.6 | Comparison of the proposed algorithm <i>HOPLP – MUL</i> with baseline algorithms in terms of AUC layer-wise on three datasets and five <i>Ratio</i> values for testing to total edges percentage (contd..) | 93 |
| 4.7 | Comparison of the proposed algorithm <i>HOPLP – MUL</i> with baseline algorithms in terms of F1 Score layer-wise on three datasets and five <i>Ratio</i> values for testing to total edges percentage | 94 |
| 4.8 | Comparison of the proposed algorithm <i>HOPLP – MUL</i> with baseline algorithms in terms of F1 Score layer-wise on three datasets and five <i>Ratio</i> values for testing to total edges percentage (contd..) | 95 |
| 4.9 | Comparison of the proposed algorithm <i>HOPLP – MUL</i> with baseline algorithms in terms of Balanced Accuracy Score layer-wise on three datasets and five <i>Ratio</i> values for testing to total edges percentage | 96 |
| 4.10 | Comparison of the proposed algorithm <i>HOPLP – MUL</i> with baseline algorithms in terms of Balanced Accuracy Score layer-wise on three datasets and five <i>Ratio</i> values for testing to total edges percentage (contd..) | 97 |

| | | |
|-----|---|-----|
| 5.1 | Comparison of the proposed algorithm <i>MNERLP – MUL</i> with baseline algorithms in terms of AUC on four datasets and five <i>Ratio</i> values for testing to total edges percentage | 118 |
| 5.2 | Comparison of the proposed algorithm <i>MNERLP – MUL</i> with baseline algorithms in terms of Balanced Accuracy score on four datasets and five <i>Ratio</i> values | 119 |
| 5.3 | Comparison of the proposed algorithm <i>MNERLP – MUL</i> with baseline algorithms in terms of F1 score on four datasets and five <i>Ratio</i> values for testing to total edges percentage | 120 |
| 5.4 | Comparison of the proposed algorithm <i>MNERLP – MUL</i> with baseline algorithms in terms of AUC layer-wise on four datasets and five <i>Ratio</i> values for testing to total edges percentage | 122 |
| 5.5 | Comparison of the proposed algorithm <i>MNERLP – MUL</i> with baseline algorithms in terms of AUC layer-wise on four datasets and five <i>Ratio</i> values for testing to total edges percentage (contd.) | 123 |
| 5.6 | Comparison of the proposed algorithm <i>MNERLP – MUL</i> with baseline algorithms in terms of Balanced Accuracy Score layer-wise on four datasets and five <i>Ratio</i> values for testing to total edges percentage | 124 |
| 5.7 | Comparison of the proposed algorithm <i>MNERLP – MUL</i> with baseline algorithms in terms of Balanced Accuracy Score layer-wise on four datasets and five <i>Ratio</i> values for testing to total edges percentage (contd.) | 125 |
| 5.8 | Comparison of the proposed algorithm <i>MNERLP – MUL</i> with baseline algorithms in terms of F1 Score layer-wise on four datasets and five <i>Ratio</i> values for testing to total edges percentage | 126 |
| 5.9 | Comparison of the proposed algorithm <i>MNERLP – MUL</i> with baseline algorithms in terms of F1 Score layer-wise on four datasets and five <i>Ratio</i> values for testing to total edges percentage (contd.) | 127 |

| | | |
|------|--|-----|
| 6.1 | The computation of likelihood score of (x,y) under $CLP - MUL$ for Example Fig.6.2 | 141 |
| 6.2 | Comparison of the proposed algorithm $CLP - MUL$ with baseline algorithms in terms of AUC | 148 |
| 6.3 | Comparison of the proposed algorithm $CLP - MUL$ with baseline algorithms in terms of F1 Score | 149 |
| 6.4 | Comparison of the proposed algorithm $CLP - MUL$ with baseline algorithms in terms of Balanced Accuracy Score | 150 |
| 6.5 | Comparison of the proposed algorithm $CLP - MUL$ with baseline algorithms in terms of AUC layer-wise | 152 |
| 6.6 | Comparison of the proposed algorithm $CLP - MUL$ with baseline algorithms in terms of AUC layer-wise (contd..) | 153 |
| 6.7 | Comparison of the proposed algorithm $CLP - MUL$ with baseline algorithms in terms of F1 Score layer-wise | 154 |
| 6.8 | Comparison of the proposed algorithm $CLP - MUL$ with baseline algorithms in terms of F1 Score layer-wise (contd..) | 155 |
| 6.9 | Comparison of the proposed algorithm $CLP - MUL$ with baseline algorithms in terms of Balanced Accuracy Score layer-wise | 156 |
| 6.10 | Comparison of the proposed algorithm $CLP - MUL$ with baseline algorithms in terms of Balanced Accuracy Score layer-wise | 157 |

Abbreviations

| | |
|-------------------|---|
| ELP | Ego-based Link Prediction |
| HOPLP-MUL | Higher Order Path-based Link Prediction for Multiplex networks |
| MNERLP-MUL | Merged Node and Edge Relevance-based Link Prediction for Multiplex networks |
| CLP-MUL | Community-based Link Prediction for Multiplex networks |
| SHOPI | Link Prediction in Complex Networks based on Significance of Higher-Order Path Index |
| CLP-ID | Clustering-based Link Prediction using Information Diffusion |
| AUC | Area Under the Receiver Operating Characteristic Curve |
| AUPR | Area Under the Precision-Recall Curve |
| CN | Common Neighbors index |
| JC | Jaccard Coefficient index |
| AA | Adamic/Adar index |
| PA | Preferential Attachment index |
| RA | Resource Allocation index |
| SP | Shortest Path index |
| COSP | Cosine+ index |
| MFI | Matrix Forest index |
| ACT | Average Commute Time index |
| LPI | Local Path index |

| | |
|-------------------|--|
| L3 | Path of length 3 index |
| CCLP | Clustering Coefficient-based Link Prediction |
| NLC | Node and Link Clustering coefficient |
| CAR | Cannistraci-Alanis-Ravasi-based Common Neighbors index |
| LNBCN | Local Naive Bayes-based Common Neighbors index |
| N2V | Node2Vec |
| SOIDP | Second Order Iterative Degree Penlaty |
| CN-WT | Common Neighbors index for Weighted graphs |
| JC-WT | Jaccard Coefficient index for Weighted graphs |
| AA-WT | Adamic-Adar index for Weighted graphs |
| PA-WT | Preferential Attachment index for Weighted graphs |
| RA-WT | Resource Allocation index for Weighted graphs |
| CC-WT | Clustering Coefficient-based index for Weighted graphs |
| LocalIP-WT | Local Path-based indexfor Weighted graphs |
| NSILR-MUL | Node Similarity Index based on Layer Relevance for Multiplex networks |
| MADMLP-MUL | Multiple Attribute based Decision Making for Multiplex networks |

Symbols

| | |
|-----------------|--|
| $G(V, E)$ | General graph G with V nodes and E edges |
| $N(i)$ | Set of directly connected neighbor of node i |
| A_M | Adjacency matrix of summarized multiplex graph without layer weight |
| $\psi^a(x)$ | Ego region set of node x of length a |
| $\psi(u, v)$ | Ego strength of existing edge between nodes u & v |
| $\gamma(x, y)$ | Node-based feature set used to calculate likelihood of non existing edge between nodes x & y |
| $G^j(V, E^j)$ | j -th layer of multiplex graph G^j with V nodes and E^j edges |
| $a_{x,y}^j$ | Edge between nodes x & y in j th layer of multiplex graph G^j |
| $\ V\ $ | Number of nodes in graph |
| $\ V\ * \ V\ $ | Number of total possible edges in graph |
| l | Current length of path considered for influence propagation |
| l_{max} | Maximum length of path considered for influence propagation |
| A_{HOPLP} | Adjacency matrix of summarized multiplex graph with layer weightage |
| G_{HOPLP} | Summarized weighted graph created from A_{HOPLP} |
| $CZ(j)$ | Compression constant for j -th layer |
| $DCZ(j)$ | Decompression constant for j -th layer |
| $IS(n_1, n_2)$ | Combined initial significance of all paths between nodes n_1 & n_2 for influence propagation |
| ψ^l | Dampening factor for l -length path |

| | |
|------------------------|---|
| $LI(n_1, n_2)$ | Likelihood of link between nodes n_1 & n_2 for summarized graph |
| $LI_j(n_1, n_2)$ | Likelihood of link between nodes n_1 & n_2 for j -th layer of multiplex graph |
| $score_{\ v\ * \ V\ }$ | Matrix with combined current influence between all node pairs |
| $prior_{\ v\ * \ V\ }$ | Matrix with combined current influence between all node pairs for previous shorter path |
| $CC(x)$ | Closeness Centrality of node x |
| $BC(x)$ | Betweenness Centrality of node x |
| $HC(x)$ | Harmonic Centrality of node x |
| $dist(x, a)$ | Shortest distance between nodes x & a |
| $\gamma(a, b)$ | Number of shortest paths between nodes a & b |
| $\gamma(a, b x)$ | Number of shortest paths between nodes a & b with x as intermediate node |
| A_M | Adjacency matrix of summarized multiplex graph without layer weight |
| A_{MNERLP} | Adjacency matrix of summarized multiplex graph |
| G_{MNERLP} | Summarized weighted graph created from A_{MNERLP} |
| $P(j)$ | Packing constant for j -th layer |
| $UP(j)$ | Unpacking constant for j -th layer |
| $ER(a, b)$ | Total relevance of existing edge between nodes a & b |
| $NR(a)$ | Total relevance of node a |
| $MR_{CN}(a, b, c)$ | Merged node and edge relevance of common neighbor b of nodes a & c |
| $MR_{CNV}(d)$ | Merged node and edge relevance of all nodes in the vicinity (directly connected) of node d |
| α | Contribution weightage of edge relevance |
| β | Contribution weightage of node relevance |
| $C_L(x)$ | Community label of node x |
| $S_I(C_L)$ | Stabilization index for community label C_L |

| | |
|------------|---|
| $I_I(x,y)$ | Individual impact of node x on node y |
| $C_I(x,y)$ | Collective impact of node x on node y based on community labels |

