Enhanced Link Prediction using Aggregation for Edge Relevance Quantification

Thesis submitted in partial fulfillment

for the Award of Degree

DOCTOR OF PHILOSOPHY

by

SHIVANSH MISHRA

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY

(BANARAS HINDU UNIVERSITY),

VARANASI-221 005

Roll No: 18071005 2023

Chapter 7

Conclusion and Future Directions

In this final chapter major contributions of the thesis are highlighted along with a discussion of future directions which are possible to extend this research.

7.1 Conclusion

Link prediction in networks remains an area of ever-growing research. This thesis has exploited local and global information to create quasi-local similarity-based link prediction approaches that apply to single-layered and multiplex networks. All the proposed algorithms in this thesis depend on edge relevance quantification using different aggregation procedures. In simple networks, the aggregation of local regions is used to estimate edge relevance. In contrast, in multiplex networks, this aggregation and relevant layer-based de-aggregation are used to make layer-wise link prediction a connected process. The edge relevance is based on how many layers the edge exists in and how relevant are the layers themselves in contrast to the aggregated graph structure. The main objective of the thesis was to use edge relevance in improving link prediction in simple and multiplex networks. The proposed works attempt to achieve a trade-off between local and global approaches such that widely applicable quasi-local similarity-based are created which are also computationally efficient.

In chapter 3, the objective of calculating efficient edge relevance to extend local similarity-based link prediction algorithms into the quasi-local domain is addressed. In this chapter, an Ego-based Link Prediction algorithm (*ELP*) is proposed and developed, which uses an ego-based link strength estimation perspective to predict target links. Classical algorithms do not consider the cumulative effect of node-based strength propagation on edges to predict target links, but that is the algorithm's specialty. Experimental results demonstrate that *ELP* performs exceptionally well in the Accuracy metric, a combined representation of the prediction performance of both existent and non-existent edges. For other metrics, i.e., AUPR and AUC, it can be observed that *ELP*'s performance is better on datasets with an average degree significantly more than 10. This makes the algorithm more suitable for link prediction of networks with the magnitude of edges is much larger than nodes.

In Chapters 4,5, and 6, the objective of employing layer aggregation as a method for simplifying link prediction in multiplex networks is addressed. In chapter 4, an algorithm was proposed and developed called Higher Order Path-based Link Prediction for Multiplex networks ($HOPLP-MUL$). This algorithm describes a unique technique for link prediction in multiplex networks based on the relevance of higher-order pathways and layer fusion. This method has sought to anticipate linkages by including more information about nodes (considerably larger zones of influence) and applying appropriate damping and layer fusion procedures on connecting paths between nodes. The density-based proposed parameters and the modified initial significance play an essential role in the $HOPLP-MUL$ method. The findings reveal that localized neighborhood-based algorithms have a relatively limited picture of the routes connecting nodes, resulting in reduced accuracy. This fact has been capitalized on in this proposal. The solution beats existing link prediction algorithms for link prediction in multiplex networks.

In chapter 5, an algorithm was proposed and developed called Merged Node and Edge Relevance based Link Prediction in Multiplex networks ($MNERLP-MUL$). This algorithm is based on merging node and edge relevance to take both local and global information into account. The proposal aimed to predict links using more information between nodes (quasi-local approach) and to better predict links in specific layers from a summarized weighted graph. The results demonstrate that local neighborhood-based algorithms take a very restrained view of overall network information to predict edges between nodes, resulting in lower accuracy. This fact has been improved upon. The variation of weightage of both edge and node relevance for link prediction has also been explored. Another characteristic is that only one round of link prediction should be performed (non-layer specific). Layer-specific link likelihoods can be calculated with just a simple multiplication with an unpacking constant.

In chapter 6, an algorithm was proposed and developed called Community-based Link Prediction on Multiplex networks ($CLP-MUL$). The proposed algorithm predicts links that are not specific to a particular layer but are based on communities detected using the summarized information of all layers. The proposed approach for link prediction considers these communities to stretch across layers even if the edge structure of a particular layer may not totally agree with it. The experiments were performed on six real-world datasets, and the results indicate that the argument was justified for datasets with low average shortest path length and relatively higher number of training edges.

7.2 Future directions of work

The following future directions are proposed for improving upon the individual novel algorithms proposed in this thesis -

- **ELP -** The *ELP* algorithm, at its core, considers an ego to be the region of influence of a node. This influence is combined for all nodes on a single edge to estimate the edge's real significance. Other formulation models of the region of influence can be used, for example, an information diffusion-based perspective. Also, variable-sized regions of influence based on node properties can be explored. Also, since the summation of node influences is the algorithm's basic proposition, dynamic networks might be a good fit for this algorithm.

- **HOPLP-MUL -** In the future, $HOPLP-MUL$ can be extended using random walks to improve its complexity. This is because, for a sparse enough network and a sufficient number of walks, random walks can give us the same exhaustive picture of a node's neighborhood as the deterministic approach used in this work. Also, this approach can be extended by incorporating a component of global node properties which can help improve the accuracy of link prediction.

- **MNERLP-MUL -** In the future, $MNERLP-MUL$ can be extended using an overall better edge relevance calculation as it is evident from the parameters fixed ($\alpha = 0.2 \& \beta = 1.0$) that the relative contribution of node and edge relevance is skewed in favor of node relevance. One possible strategy for this can be using graph properties to define a variable $\alpha \& \beta$ instead of a fixed parameter approach. Another would be employing better methods of edge relevance quantification such that the contribution of both edge and node relevance can be equalized.

- **CLP-MUL -** In the future, $CLP-MUL$ can be extended by identifying the importance of particular communities to different layers of the multiplex networks and then utilizing this distinguishing information to improve link prediction in combination with the simple layer density-based approach. Community detection methods can also be employed other than label propagation ones, which can give better link prediction results.

Based on the above future directions about the proposed algorithms in thesis, the following areas are suggested for researchers in which link prediction can be used -

- **Node identification -** Link prediction is essentially a graph completion task. The knowledge gleaned from these more complete graphs can be used to improve localized graph matching which can help in node labeling dis-separate layers into coherent multiplex networks.

- **Cluster identification -** Many cluster identification and community detection techniques have been developed for multiplex networks [172]. If link prediction is used as a preprocessing step for this task, the accuracy of cluster detection may be enhanced on account of more relevant information availability.

- **Epidemic response -** In the post-COVID times, the epidemic response has become a vital area of interest. Systems that track the spread of a virus against vaccinated and infection-based immunized clusters have become imperative in this decision-making process [173]. Link prediction can help in predicting future directions of epidemic spread and most effective response strategies.

- **Genetics and other living research -** Multiplex networks provide an avenue for modeling increasingly complex biological systems, which range from extremely minute protein-protein interactions within DNA to increasingly more significant cell subgroups and organ systems [174]. Since even the simplest organisms found in nature are more complex that the best machines humans can build, the identification of links in such systems help in enhancing the understanding of living beings' physiology.

- **Transportation modeling -** Multiplex and multilayer networks have been extensively used as an abstraction of the resource-constrained transportation modeling problem for living beings [175]. When this problem is extended into human-related transportation systems, it is converted into a traffic optimization problem [176–179], which is of paramount interest for current researchers as many big cities in the world are facing crippling traffic jams due to automobile and population increases. Even rail and air transport systems can be modeled using multiplex networks where new link identification can help in better future planning, network congestion identification and improved redundancy.