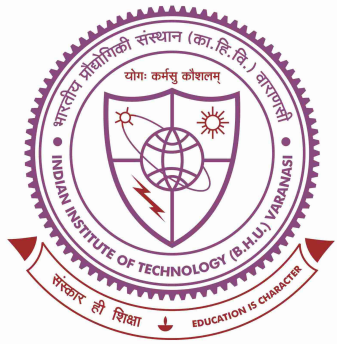
Link Prediction in Dynamic Networks using Feature and Quantum based Machine Learning Techniques



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

Chapter 8

Conclusion and future directions

In this final chapter major contributions of the thesis are highlighted along with the discussion of future directions.

8.1 Summary of Contributions

Link prediction in the networks has remained utmost important in social network analysis. In this thesis, we exploited several topological or structural information of networks to compute similarity scores of node-pairs. We encode these information as discriminating features for link prediction framework. In order to give the best feasible solution to the link prediction problem, these individual features and their combinations were examined with various machine learning models. Main contributions of the thesis are listed below.

- **Feature fusion-based link prediction:** The three widely used similarity-based indices, Local (L), Global (G), and Quasi-local (Q), are explored in different combinations (L, G, Q, LG, LQ, GQ, LGQ) for rich feature set generation that can be used with various machine learning techniques for link prediction. The addition of global and quasi-local similarity scores shows significant changes to link

prediction performance. After combining, these features represent different properties of the graph ranging from local neighborhoods to the full structure of the graph. The patterns in these properties help us in improving the performance of link prediction in dynamic graphs by a decent margin. We have also tested the performance of these features with different machine learning models (NN and XGB) to further enhance the understanding of the pairing of different types of feature sets with different models. Experimental results show on six datasets that Neural Network based prediction model shows the best performance with feature sets having a low number of features and has a preference for quasi-local similarity indices, while XGBoost doesn't have any such preference o specific datasets and works best with LGQ based feature set.

- **PWAF:** The Path Weight-Based Aggregation Feature (PWAF), a new feature, is proposed. In addition to the recommended Path Weight-Based Aggregation Feature, several topological properties of the networks (Local, Global, and Quasi-local), as well as Clustering Coefficient based features, are taken into consideration for feature generation. The Level-2 node clustering coefficient (CCLP2) is one of the features used to improve prediction. For link prediction, many machine learning models are used to make predictions using this rich feature set, including Neural Network (NN), Logistic Regression (LR), XGBoost (XGB), Random Forest Classifier (RFC), and Linear Discriminant Analysis (LDA). The experiments are carried out on seven different well-known dynamic network data sets in terms of five performance evaluation metrics. Among all algorithms and state-of-the-art approaches, PWAF-RFC is the top performer. In addition, PWAF-XGB also provides superior performance among individual features and state-of-the-art methods.

- **CFLP:** It incorporates features that perform edge scoring individual snapshots and the whole dynamic graph through the feature set. The proposed feature CFLP is used for estimating edge behavior throughout the entire dynamic network irrespective of the particular snapshot under consideration. Four similarity indices

for estimating edge behavior on individual snapshots: local similarity-based, global similarity-based, quasi-local similarity-based, and clustering coefficient-based similarity. In order to discover the optimal combination of minimally optimized features for link prediction, feature selection is also applied to fourteen different snapshot-based features. This combined feature set, which incorporates various categories of similarity, is tested with different machine learning models and compared with individual features and state-of-the-art algorithms to verify the improved performance of our approach. Experiments are conducted on seven real-world datasets, and the proposed feature set provides the best performance on XGB and RFC models with five evaluation metrics.

- **Community Enhanced Link Prediction:** Developed a framework to generate community information-based link prediction features. These features enhance link prediction performance, typically conducted only with local, global, and quasi-local similarity-based features. Using feature relevance scoring, we demonstrated the superiority of community features compared to some standard topological link prediction features. Finally, we provide an optimized feature set version of a combination of our community-based features with traditional link prediction-based features, COMMLP-DYN. This feature set performs better than other state-of-the-art algorithms for link prediction on dynamic networks in a snapshot-based setting. The experiments were conducted on six real-world datasets, three ratio training values to total edges, and three different performance metrics.

- **PQKLP:** Proposed a novel strategy for addressing the link prediction problem using Projected Quantum Kernel (PQK) enhanced machine learning models, which utilize both local and global information for feature generation. The goal of our research is to create a quantum-assisted feature-based new approach for link prediction that integrates Projected Quantum Kernel (PQK) with machine learning models to increase prediction performance. By using PQK, we enhanced our data using high-dimensional Hilbert spaces to achieve improved link prediction.

Quantum models look at data in high-dimensional Hilbert spaces, to which, in other cases, we can only have access through inner products revealed by measurements because they have a mathematical structure that is similar to that of quantum mechanics. The experimental results on five well-known dynamic datasets demonstrate that the new quantum-assisted feature-based technique outperforms the corresponding machine learning models in some cases, especially for PQKLP-NN and PQKLP-RFC.

8.2 Scope for Future Work

Although the literature has explored a number of link prediction methods, the problem is still disputed. There are still a number of issues that need to be addressed, such as which structural properties work best with certain approaches and how to handle the complexity of the network. Handling imbalanced datasets in the context of link prediction may be another challenge because the majority of real-world networks are relatively sparse, with very few positive occurrences compared to negative instances. There are few published research on knowledge graphs, attributed graphs, multiplex, and multi-layer networks; these topics can be further addressed in the future.

In this section, several new avenues of research that have been opened up by this thesis are mentioned below.

- **LGQ:** New feature sets with more similarity scores that target different properties of social networks can be developed. The performance of this LGQ model for link prediction on heterogeneous networks is also possible for extending this work.
- **PWAF:** Estimating such features using information derived from dynamic community detection models can be explored in this approach. It can also be extended for use on heterogeneous networks where nodes and edges are classified into different classes.

- **CFLP:** Snapshot-impartial features that make use of combination of three major types of graph properties (degree, path, clustering coefficient) can be explored. Dimensionality reduction methods other than feature selection can also be experimented with.

- **Community enhance link prediction:** This framework can be modified to make use of extended community detection methods that have been explicitly defined for dynamic graphs. Also, a new formulation can be researched which considers overlapping community information rather than non-overlapping ways whose performance was discussed in this work.

- **PQKLP:** More experiments can be conducted with expanded feature sets and feature ranking can be used to optimize on their performance in this framework. This work can also be expanded to the domain of other types of networks, such as multiplex and attributed ones.