Structural Similarity-based Link Prediction in Complex Networks

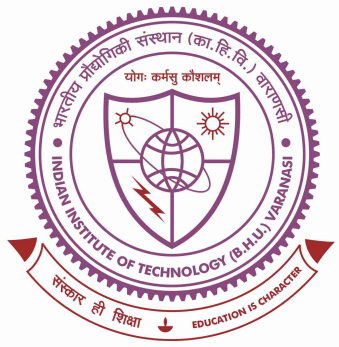
Thesis submitted in partial fulfillment

for the Award of Degree

DOCTOR OF PHILOSOPHY

by

AJAY KUMAR



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY

(BANARAS HINDU UNIVERSITY),

VARANASI-221 005

Roll No: 16071005

2021

Chapter 6

Conclusion and future directions

6.1 Conclusion

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. Based on these features, we proposed two novel similarity methods of link prediction where one of the method use clustering coefficient while the other use paths information.

In Chapter 3 as a survey, we have gone through several link prediction methods broadly classified into similarity-based, probabilistic models, dimensionality reduction-based, entropy-based, and clustering-based. We have also reviewed some recent approaches, including fuzzy models and link prediction in bipartite networks. The experiment of similarity-based approaches on seven network datasets has been conducted and evaluated on four well-known measures. We observed that local and quasi-local approaches perform well, usually. Global approaches are mostly based on exploring paths which are complex to compute and increase the noise in the networks. The running time of these methods with their algorithmic complexity in big 'O' notation has been reported in this survey.

In Chapter 4, we empirically investigated (using 11 network datasets) the effect of traditional clustering information for several existing techniques of link prediction. It is observed that this information shows better accuracy results compared to common neighbor based methods. With this observation and the intuition that more structural information of networks enhance the accuracy, we extended the notion of node clustering coefficient to higher level (i.e. level-2). For this, notion of level-2 common neighbor is given in the chapter. Once these level-2 common neighbors are identified for a given node-pair, clustering coefficient of these nodes are computed and summed to finalize the similarity score of the node-pair. Experimental results reveal its superiority over the existing methods.

In Chapter 5, we explore other topological features in link prediction framework. Here, path features with different lengths are analyzed. Path of longer length can be utilized to extract more structural information of a network. The proposed work in this chapter finds similarity score of a node-pair which is based on the resource allocation process in networks. This concept states that the similarity between the two nodes is amount of information received by receiver node that is sent by the sender node. To maximize this information, we minimize the information leaks in the paths. Less information leak derives more similarity between nodes. We encode these information as discriminating feature for link prediction framework and proposed a novel method. Effects of different path lengths as well as penalization factor for different methods are presented. Experimental results show its accuracy enhancement on datasets from diverse areas.

6.2 Future directions

Although several link prediction methods have been explored in the literature, it is still an open research problem. Several problems are yet to be explored, for example, which structural properties perform better on each technique, also how to deal with the large size of the network. Can we devise an approach to predict missing links where

strengths/weights are changing with time? As outlier concept is useful to detect spam emails, so outliers detection may be another framework where link prediction approaches would make a fruitful contribution. Most real-world networks are highly sparse where the number of positive instances is very few compared to negative instances, so handling imbalanced datasets in the context of link prediction may be another aspect. Limited works on multiplex and multi layer networks are available in the literature; this can be more explored in the future. Today's communications comprise several recipients (more than 2) that shift our attention from a one-to-one communication to one-to-many and many-to-many. Link prediction can be useful in such scenarios.