INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
SHORT ABSTRACT OF THESIS

Name of the Student : Akash Anil

Roll Number : 146101015

Programme of Study : Ph.D.

Thesis Title: **Link Prediction in Heterogeneous Information Networks: From Network Topology to Network Embedding**

Name of Thesis Supervisor(s) : Dr. Sanasam Ranbir Singh

Thesis Submitted to the Department/ Center : Computer Science & Engineering

Date of completion of Thesis Viva-Voce Exam : 05/02/2020

Key words for description of Thesis Work : Link Prediction, Social Network Analysis, Heterogeneous Information Networks, Network Embedding

SHORT ABSTRACT

Modeling real-world systems using complex network analysis has become a popular approach in the last two decades. A complex network is loosely divided into four types of networks, namely (i) Social Network, (ii) Information Network, (iii) Technological Network, and (iv) Biological Network. However, most real-world networks can be represented as Information Network. Majority of the previous literature over information networks consider homogeneous network representation (singular types of nodes and relations) e.g., Citation network, World Wide Web, etc. However, it has recently been realized that Heterogeneous Information Network (HIN) that consists of multiple types of nodes and relations is a better representation for real-world physical systems. For example, a HIN representing a Citation network by considering node types, such as Author, Paper, Venue, etc. and their corresponding relations, captures rich semantics in comparison to the homogeneous Citation network (considering only the papers as node and citation as relation). Motivated with this, our objective is to leverage Heterogeneous Information Network representation in modeling evolution of a given system by solving link prediction problem. In particular, the major contributions of this thesis towards link prediction can be divided into three types of approaches; (i) topology-based, (ii) graph kernel-based and (iii) network embedding-based. For topology-based methods, we adapt the state-of-the-art common neighbor-based local similarity measures to heterogeneous information network. For graph kernels-based methods we propose a generalized heterogeneous framework for state-of-the-art spectral graph kernels. Furthermore, in network embedding-based methods, we exploit k- hop random-walk to generate node neighborhood for training the model. From previous studies, it is evident that majority of the complex networks are susceptible to the exogenous information (e.g., news and social media) apart from endogenous information (e.g., network characteristics such as clustering coefficients, degree distribution, etc.). Therefore, we study the effects of exogenous information such as news media, temporal dynamics of the underlying network on the proposed topology-based heterogeneous similarity measures. We observe that incorporating exogenous information helps in boosting performance of the link prediction. Further, majority of the studies on link prediction using network embedding are based on meta paths, we critically analyze the efficacy of meta path-based methods over link prediction and node classification tasks. We observe that heterogeneous network embedding cannot be generalized and meta path-based embeddings are task-specific. As most of the heterogeneous information network having different number of instances for different types of nodes and relations, class imbalance is inherent in such networks. Therefore, this thesis further studies the effects of class imbalance in heterogeneous network embedding. It is observed that selecting appropriate node types along with addressing class imbalance in heterogeneous information network is an important pre-requisite for efficient network embedding.