Thesis Title: Structure-Aware Network Representation Learning on Graphs

Short Abstract

Graphs provide a powerful way to model real-world scenarios. Entities in real-world data can be modeled as nodes in a graph. And, various kinds of interactions among the entities are modeled as edges in such a graph. It is a ubiquitous data structure to represent linked data from diverse domains --- social, technological, biological, financial, transports, cellular networks, recommender systems, and many more. Several classes of graphs such as homogeneous, heterogeneous, multiplex graphs exist that cater to the need for modeling the specific nature of node interactions in real-world data. Nevertheless, modeling real-world data as graphs often suffers from challenges such as noise, incomplete and unobserved information, sparsity, heterogeneity, structural diversities, lack of annotations, the existence of imbalanced graph components, inter-dependencies of graph components, etc. Mining useful information and obtaining inference from the graphs are of utmost importance and have immense applicabilities. Traditional Machine Learning algorithms require useful features from the graphs as input to obtain insightful inferences. Earlier approaches to feature engineering often require user-defined heuristics, the intervention of domain experts, and critical resources. Recently, Graph Embedding, aka Network Representation Learning (NRL), has become a very popular means of automatically extracting useful latent features from graphs, especially for large graphs. The NRL methods aim to learn a mapping function to project high-dimensional non-Euclidean graph data (for representing nodes, edges, paths, subgraphs, or the entire graph) into a low dimensional latent embedding space optimally without compromising the underlying structural properties of the original graph.

In a trivial setup, NRL aims at incorporating local neighborhood contexts, aka microscopic views surrounding graph elements of interest. The microscopic views are often inadequate in learning discriminative features for various downstream tasks. As observed in recent studies, learning higher-order macroscopic views (global structure) can improve the discriminative capacity of the features for various applications. However, based on the complexities and scale of the underlying network, capturing the macroscopic view is a non-trivial task. This dissertation carefully examines existing research gaps while incorporating macroscopic views and proposes three novel network embedding methods for homogeneous, heterogeneous, and multiplex graphs.

Past research efforts in learning structures in homogeneous graphs primarily focus on either capturing k-hop local neighborhood contexts of nodes or jointly learning communities to enrich node embeddings. All the community enforcing models use unsupervised clustering criteria based on either network-only node proximities or embedding-based node proximities. To address this, the thesis first investigates incorporating supervised non-network node proximity measures to group nodes in homogeneous graphs. The framework unifies ways to include supervision knowledge for enriched node embedding learning. Robust node classification and clustering performance are obtained even in challenging experiment setups, with varying ratios of class labels and different node-sampling strategies.

Next, this thesis considers improving the InfoMax based learning strategy as a useful means to incorporate global graph structures into node embeddings for multiplex graphs. InfoMax based learning provides a scalable way to incorporate both local and global node representations via maximizing Mutual Information (MI) between them. Nevertheless, in a typical setup, it uses a common global graph summary for all the local node embeddings, thereby encoding a lot of noisy and trivial information. The thesis proposes a novel way to contextualize global graph summaries for each node to encode non-trivial personalized graph summaries in node embeddings. The effectiveness of the proposed framework is verified with several downstream tasks, such as node classification, clustering, and similarity-search.

Finally, this thesis considers incorporating various structural contexts at multiple granularities between the nodes for improving link prediction performance in heterogeneous graphs. Very few research efforts have been made to understand various structural cues that exist for link prediction in heterogeneous graphs. Also, no NRL study has investigated the roles that communities of the end nodes play in predicting links between the nodes. To address this, this dissertation proposes a novel, first-of-its-kind community view of the edges in a graph. The proposed framework considers relational paths between the nodes and their communities apart from the popularly used common subgraph contexts. It also proposes a fine-grained attention mechanism to combine all the candidate contexts judiciously for link prediction. The framework outperforms the most recent benchmark heterogeneous link prediction method by a huge margin. Visualizing attention weights of candidate structural contexts establishes their usefulness and complementarity in aiding link prediction at various challenging scenarios. This dissertation shows that learning structure-aware network representations facilitates learning of enriched target graph-component embeddings that can benefit various downstream ML tasks.