Department of Systems Science & Industrial Engineering

Learning Large-Scale Network Representations

Dissertation Defense

Saba A. Al-Sayouri, PhD Candidate

Thursday, May 9$^{\text{th}}$, 1-3pm, EB-R3

Abstract

Real-world networks, such as social networks, biological networks, citation networks, language networks, terrorism networks, among others, describe simple and sophisticated relationships connecting various entities together, where such relationships are encoded in a graph format. Modeling complex connectivity patterns can help uncover and understand the reciprocal intrinsic interactions among network's entities in a systematic manner. However, since real-world networks come in very large scales (millions of nodes and billions of edges), mining these networks is impeded by the networks' curse of dimensionality. In response to that, various solutions have been proposed including feature engineering and feature learning (a.k.a. representation/embedding learning).

Contrary to network feature engineering methods, network representation learning (NRL) techniques are more efficient and generalizable, while performance-analogous. Therefore, it has received sustained attention from the research community. Following the contemporary surge of interest, a plethora of NRL techniques have been developed in the literature. They are abstracted into two categories: (1) Conventional methods, such as matrix factorization-based methods ; and (2) Recent methods that are divided into three streams: (a) Random walk-based, (b) Edge modeling-based, and (c) Deep learning-based techniques. However, the developed methods in the literature share a set of shortcomings, which are: (1) Sparsity and Effectiveness; (2) Efficiency; (3) Robustness; (4) Joint learning; and (5) Interpretability. To this end, this study aims to address the limitations of the existing NRL methods. It proposes two NRL algorithms:

1. **SURREAL**: It employs the analogy between networks and corpora at one end, and networks and electrical circuits on the other, to address the effectiveness, sparsity, and robustness issues imposed by the existing NRL methods. Indicatively, SURREAL outperforms state-of-the-art techniques by up to 37% with respect to Micro-F1 score on various multi-label classification problems. Further, in contrast to state-of-the-art approaches, SURREAL, being deterministic, is stable and thus can generalize to single and multi-graph tasks.

2. **t-PINE**: It addresses a set of NRL fundamental challenges, such as effectiveness, sparsity, interpretability, and joint learning, using the notion of tensor decomposition. t-PINE is capable to learn more useful, highly predictable, and gracefully interpretable representations. Remarkably, t-PINE outperforms baseline methods by up to 351.5% with respect to Micro-F1 score on different multi-label classification problems, while it has high visualization and interpretability utility.

Network representation learning and tensor decomposition techniques offer great benefit in the healthcare domain. This is rooted in the deep learning capabilities such frameworks provide to model the highly non-linear sophisticated relationships in healthcare data.