



Proposal Defense

Doctor of Philosophy in Computer Science

“Causal inference using deep-learning-based variable selection” by **Zhenjiang Fan**

Date: June 24, 2021

Time: 12:00pm – 1:00pm

Place: https://pitt.co1.qualtrics.com/jfe/form/SV_dj1DOviMx3atWXY

Committee:

- Taieb Znati, Professor, Department of Computer Science, School of Computing and Information
- Hyun Jung Park, Assistant Professor, Human Genetics, School of Public Health
- Seong Jae Hwang, Assistant Professor, Department of Computer Science, School of Computing and Information
- Xulong Tang, Assistant Professor, Department of Computer Science, School of Computing and Information
- Adriana Ivanona Kovashka, Assistant Professor, Department of Computer Science, School of Computing and Information
- Heng Huang, John A. Jurenko Endowed Professor, Department of Electrical and Computer Engineering

Abstract: Causal structure learning plays a critical role in biomedical studies because the underlying causal relations between variables may help identify driver genes and therapeutic agents. Although several algorithms have been developed to learn the causal structure, most of them do not identify causal relations in the interactions that are through multiple mediating layers (indirect associations). Since biological variables interact through multiple regulatory layers and have indirect associations in complex biological systems, we hypothesize that considering such layers enables us to uncover causal relations that were not previously identified. Based on this hypothesis, we propose a method that brings in two main advantages. First, it enables us to learn causalities in the associations through multiple mediating layers (indirect associations). Second, unlike regular deep-learning methods, it can estimate the strength of the associations. This proposed work uses the concept of knockoff control and deep neural networks (DNNs) to identify indirect (nonlinear) associations; it also uses a mixed graphical model (MGM) to identify direct (linear) associations; then it identifies causal directions using a scoring criterion. The proposed work aims to handle the data that contains a response variable (here, we call this type of data “labeled data” since each of the samples is categorized or related to a response variable or label) as well as the data that have no response variables (here, we call this type of data “correlation data” as we aim to study relationships among the data variables). Another goal of the proposed work is to process mixed data in which both continuous and discrete variables are presented. As discussed above, the proposed method is also able to discover both linear and nonlinear associations using DNNs and MGMs. We are planning to test it using both simulated data and real-world datasets.