Dissertation Defense
Doctor of Philosophy in Computer Science

“Relational Machine Learning Algorithms” by Alireza Samadian

Date: July 7, 2021
Time: 10:00 AM – 11:00 AM
Place: RSVP to receive the Zoom information

Committee:

- Dr. Kirk Pruhs (advisor), Professor, School of Computing and Information, University of Pittsburgh
- Dr. Panos Chrysanthis, Professor, School of Computing and Information, University of Pittsburgh
- Dr. Adriana Kovashka, Assistant Professor, School of Computing and Information, University of Pittsburgh
- Dr. Benjamin Moseley, Associate Professor, Tepper School of Business, Carnegie Mellon University

Abstract:
The majority of learning tasks faced by data scientists involve relational data, yet most standard algorithms for standard learning problems are not designed to accept relational data as input. The standard practice to address this issue is to join the relational data to create the type of geometric input that standard learning algorithms expect. Unfortunately, this standard practice has exponential worst-case time and space complexity. This leads us to consider what we call the Relational Learning Question: “Which standard learning algorithms can be efficiently implemented on relational data, and for those that cannot, is there an alternative algorithm that can be efficiently implemented on relational data and that has similar performance guarantees to the standard algorithm?”

In this dissertation, we address the relational learning question for the well-known problems of support vector machine (SVM), logistic regression, and k-means clustering. First, we design an efficient relational algorithm for regularized linear SVM and logistic regression using sampling methods. Then for linear SVM, we show how to implement a variation of gradient descent that provides a nearly optimal approximation guarantee for stable instances. For the k-means problem, we show that the k-means++ algorithm can be efficiently implemented on relational data, and that a slight variation of adaptive k-means algorithm can be efficiently implemented on relational data while maintaining a constant approximation guarantee. On the way to developing these algorithms, we give an efficient approximation algorithm for certain sum-product queries with additive inequalities that commonly arise.