**Thesis:** Efficient Spacial Data Partitioning For Distributed kNN Joins

**Abstract:** Data collection is one of the most common practices in today’s world. The data collection rate has rapidly increased over the past decade and is not showing any signs of decline. Data sources are many; the Internet of Things devices, mobile gadgets, social media posts, connected cars, and web servers constantly report on their users’ interactions and habits. Much of the collected data is spatial data which contains attributes that denote the physical origin of the data. As a result of the tremendous growth in data collection, higher demand for new techniques emerged to efficiently process and extract valuable insights in a relatively acceptable time frame. The current standard approach to large-scale data analysis uses distributed parallel processing systems like Apache Hadoop and Apache Spark. However, these systems are designed for general-purpose parallel processing and require an additional layer to recognize and efficiently process spatial datasets. Motivated by its many applications, we examine the several challenges facing spatial data partitioning and processing and propose solutions customized for each task. We detail our techniques for building spatial partitioners over large datasets for use with spatial queries like map-matching and kNN spatial join. Additionally, we present an accuracy benchmarking framework for comparing and classifying the results of two input files based on specific criteria. Our proposed work targets batch processing of large spatial datasets, including structured, unstructured, and semi-structured datasets.

**Committee:**
- Professor Huy T. Vo, Mentor, The City College
- Professor Feng Gu, College Of Staten Island
- Professor Kaliappa Ravindran, The City College