COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF MICHIGAN

DISSERTATION DEFENSE

Yin Lin

Exploring Intersectionality in Responsible Data Management

Monday, November 11, 2024
10:00am – 12:00pm
3725 Beyster
Hybrid – Zoom

ABSTRACT: Big data is extensively used to train AI systems and to develop algorithms for various decision-making tasks. However, these data-driven systems and algorithms are only as reliable as the data they are built upon, which often contains human biases embedded in the pipelines or reflect historical biases inherited from the collected datasets. Without sufficient care, this "bias in, bias out" issue prevents decision-making systems from producing equitable outcomes, leading to undesirable biases, particularly against minorities frequently underrepresented in datasets and overlooked in data management processes.

My research inspects the data pipelines of such systems to enhance fairness and transparency in the use of big data. While existing studies on system fairness typically focus on a limited number of predefined groups along a single identity axis (e.g., race or gender), my work examines the intersectional effects of multiple protected attributes. In this talk, I will present a suite of techniques aimed at both modeling unfairness and designing efficient algorithms to address these challenges. First, I discuss methods for exploratory pre-training analysis and remediation, specifically considering representation bias, where certain populations are underrepresented in datasets. I develop efficient algorithms to identify underrepresented populations across multi-table databases. Additionally, I formally illustrate the connection between representation bias in training data and outcome divergence among subgroups in machine learning prediction results, proposing mitigation strategies based on dataset remediation. Second, inspired by real-world cases where news reports cherry-pick generalized conclusions that do not accurately reflect all subgroups, I propose a framework to evaluate, detect, and revise such misleading conclusions derived from aggregation. Finally, I examine row-level data lineage in data science pipelines, particularly those with non-relational operators and user-defined functions, as often seen in machine learning and analytical pipelines. I present an efficient platform that traces the lineage of erroneous or individual records at the pipeline's output, offering essential support for data debugging and integration.

CHAIR: Prof. H.V. Jagadish