Dissertation Defense
Doctor of Philosophy in Information Science

“Discovery and Interpretability of Spurious Associations in Data-Driven Decisions” by Xian Teng

Date: April 2, 2024
Time: 9:00 a.m. – 12:00 p.m.
Place: Room 522, Information Sciences Building, 135 North Bellefield Avenue, Pittsburgh, PA 15260

Committee:
- Rosta Farzan, Associate Professor, School of Computing and Information, University of Pittsburgh
- Yu-Ru Lin, Associate Professor, School of Computing and Information, University of Pittsburgh
- Peter Brusilovsky, Professor, School of Computing and Information, University of Pittsburgh
- Gregory Cooper, Professor, Department of Biomedical Informatics, University of Pittsburgh

Abstract:
Big data and machine learning tools have jointly empowered humans in making data-driven decisions. Many of them capture empirical associations that might be spurious due to confounding factors and subgroup heterogeneity. The famous Simpson’s paradox is such a phenomenon where aggregated and subgroup-level associations contradict with each other, causing confusion and decision difficulties. Existing algorithms and systems offer little support for humans – especially non-experts of causal inference and machine learning – to identify subgroups free of spuriousness, locate, reason about, and prevent pitfalls of spurious association in practice.

This dissertation introduces a novel solution to address these issues, comprising a data-driven algorithm, a human-centric workflow, and a visual analytic system. The algorithm, De-paradox Tree, uncovers subgroup patterns behind paradoxical associations by considering varying intervention probabilities and effects across subgroups. Its partition criteria are designed to balance covariates and to search for effect reversal throughout the recursive. The workflow, De-paradox Workflow, incorporates the needs of data practitioners and domain experts, empowering non-experts to effectively answer causal questions. The system, VISPUR, includes a Confounder Dashboard, a Subgroup Viewer, a Reasoning Storyboard and an interactive Decision Diagnosis panel, to help identify confounding factors, investigate diverse subgroup patterns prone to misinterpretations of causality, illustrates paradoxical phenomena and promote accountable decision-making.

Evaluations on synthetic, hybrid, and real-world datasets show that the proposed algorithm outperforms state-of-the-art baselines by building simpler trees with relevant covariates as splitting rules, deriving balanced subgroups, and disclosing nested opposites effects. Expert interviews and controlled user experiments further confirm the effectiveness of the proposed workflow and visual analytic system in helping users identify and comprehend spurious associations, as well as make accountable causal decisions.

My dissertation bridges causal theory and practical applications in observational studies, emphasizing discoverability and interpretability. This integrated solution offers practical tools applicable across diverse domains for a broad audience. It has the potential of enhancing
awareness of spurious associations, aiding reliable decision-making amid paradoxical observations and fostering knowledge discovery in data-driven research.