Mathematics, Statistics, and Computer Science **@ UIC**

Statistics and Data Science Seminar

Balancing Inferential Integrity and Disclosure Risk via Model Targeted Masking and Multiple Imputation

Bei Jiang (University of Alberta)

Abstract: There is a growing expectation that data collected by government-funded studies should be openly available to ensure research reproducibility, which also increases concerns about data privacy. A strategy to protect individuals' identity is to release multiply imputed (MI) synthetic datasets with masked sensitivity values (Rubin, 1993). However, information loss or incorrectly specified imputation models can weaken or invalidate the inferences obtained from the MI-datasets. We propose a new masking framework with a data-augmentation (DA) component and a tuning mechanism that balances protecting identity disclosure against preserving data utility. Applying it to a restricted-use Canadian Scleroderma Research Group (CSRG) dataset, we found that this DA-MI strategy achieved a 0% identity disclosure risk and preserved all inferential conclusions. It yielded 95% confidence intervals (CIs) that had overlaps of 98.5% (95.5%) on average with the CIs constructed using the full, unmasked CSRG dataset in a work-disability (interstitial lung disease) study. The CI-overlaps were lower for several other methods considered, ranging from 73.9% to 91.9% on average with the lowest value being 28.1%; such low CI-overlaps further led to some incorrect inferential conclusions. These findings indicate that the DA-MI masking framework facilitates sharing of useful research data while protecting participants' identities.

Wednesday, November 3 at 4:00 PM in Zoom

This a joint work with Adrian Raftery (University of Washington), Russel Steele (McGill University) and Naisyin Wang (University of Michigan).

Wednesday, November 3 at 4:00 PM in Zoom