JOHN HEALY, Tutte Institute for Mathematics and Computing
Exploiting distortions in clustering and dimension reduction for unstructured data

Understanding data is crucial for generating useful hypotheses and identifying various problems and biases that may be present in a dataset. Unstructured data can make this exploration challenging, but recent advances in vectorization methods have enabled the conversion of unstructured data into meaningful vector representations. Dimension reduction and clustering algorithms are powerful techniques for exploring and gaining insights from such vectorized data. However, these techniques inherently distort data—sometimes in useful ways, and sometimes problematically. In this talk, we examine the strengths, assumptions, and distortions inherent in some of the most popular clustering and dimension reduction techniques with a focus on some of the methods developed at the Tutte Institute.