Statistics Seminar

Department of Mathematical Sciences

DATE:	Thursday, Month 31, 2017
TIME:	1:15pm – 2:15pm
LOCATION:	WH 100E
SPEAKER:	Fan Yang, Binghamton University
TITLE:	Visualizing Topics with Multi-Word Expressions

Abstract

We describe a new method for visualizing topics, the distributions over terms that are automatically extracted from large text corpora using latent variable models. Our method finds significant n -grams related to a topic, which are then used to help understand and interpret the underlying distribution. Compared with the usual visualization, which simply lists the most probable topical terms, the multi-word expressions provide a better intuitive impression for what a topic is “about.” Our approach is based on a language model of arbitrary length expressions, for which we develop a new methodology based on nested permutation tests to find significant phrases. We show that this method outperforms the more standard use of chi-square and likelihood ratio tests. We illustrate the topic presentations on corpora of scientific abstracts and news articles.