Stanford University
Department of Statistics

Departmental Seminar

4:30pm, Tuesday, February 11, 2020
Sloan Mathematics Center Room 380C

Refreshments served at 4pm in Sequoia Lounge.

Speaker: Alon Kipnis, *Stanford Statistics*

Title: Two-Sample Problem for High-Dimensional Multinomials and Testing Authorship

Abstract:

The Higher Criticism (HC) test is a useful tool for detecting the global significance of multiple independent tests, especially for rare and weak effects. We adapt the HC test to a discrete two-sample setting and use it as a measure of similarity between the samples. We apply this measure to word-frequency tables and authorship attribution challenges, where the goal is to identify the author of a document using other documents whose authorship is known. The method is simple yet performs well without handcrafting and tuning. Furthermore, as an inherent side effect, the HC calculation identifies a subset of discriminating words, which allow additional interpretation of the results. Our examples include authorship in the Federalist Papers and machine-generated texts.

We take two approaches to analyze the success of our method. First, we show that, in practice, the discriminating words identified by the test have low variance across documents belonging to a corpus of homogeneous authorship. We conclude that in testing a new document against the corpus of an author, HC is mostly affected by words characteristic of that author and is relatively unaffected by topic structure. Finally, we analyze the power of the test in discriminating two multinomial distributions under rare and weak perturbations. We derive a phase transition curve for the power of the test which separates the parameter space into an area where the test is successful and an area where it fails. This phase curve is different than the phase curve in the Gaussian means model.