

# Department of Applied and Computational Mathematics and Statistics Colloquium



**Jason Klusowski**

Department of Statistics & Data Science  
Yale University

## *Counting motifs and connected components of large graphs via subgraph sampling*

Learning properties of large graphs from samples is an important problem in statistical network analysis. We revisit the problem of estimating the number of connected components in a graph of  $n$  vertices based on the subgraph sampling model, where we observe the subgraph induced by  $n$  vertices drawn uniformly at random. The key question is whether it is possible to achieve accurate estimation, i.e., vanishing normalized mean-square error, by sampling a vanishing fraction of the vertices. We show that it is possible by accessing only sublinear number of samples if the graph does not contain high-degree vertices or long induced cycles; otherwise it is impossible. We obtain optimal sample complexity bounds for several classes of graphs including forests, cliques, and more generally, chordal graphs, achieved by linear-time estimators. The methodology relies on topological identities of induced subgraph counts. They, in turn, also play a key role in proving minimax lower bounds based on constructing random instances of graphs with matching structures of small subgraphs. We will also discuss results for the neighborhood sampling model, where we additionally observe the edges between the sampled vertices and their neighbors. In this setting, we will show how to construct optimal estimators of motif counts that are adaptive to certain unknown graph parameters.

**Friday, January 12, 2018**

**4:30 PM – 5:30 PM**

**127 Hayes-Healy Center**

Colloquium Tea 4:00 PM to 4:30 PM 154 Hurley Hall