DEPARTMENT OF MATHEMATICAL SCIENCES

Seminar on Numerical Methods
Joint with Statistics Seminar

Mikhail Zaslavsky

Schlumberger-Doll Research

Title: Clustering of graph vertex subset via Krylov subspace model reduction

ABSTRACT: Clustering via graph-Laplacian spectral imbedding is ubiquitous in data science and machine learning. It provides a low dimensional parametrization of the data manifold which makes the subsequent clustering (with, say, k-means or any of its approximations) much easier. However, it becomes less efficient for large data sets due to two factors. First, computing the partial eigendecomposition of the graph-Laplacian typically requires a large Krylov subspace. Second, after the spectral imbedding is complete, the clustering is typically performed with various relaxations of k-means, which may lose robustness with respect to the initial guess, become prone to getting stuck in local minima and scale poorly in terms of computational cost for large data sets.

Normalized graph-Laplacian is intimately related to the random walk on the graph, and we will exploit this connection in our algorithms. In particular, we propose two novel algorithms for spectral clustering of a subset of the graph vertices (target subset) based on the theory of model order reduction. They rely on realizations of a reduced order model (ROM), that accurately approximates the transfer function of the random walk on the original graph for inputs and outputs restricted to the target subset. While our focus is limited to this subset, our algorithms produce its clusterization that is consistent with the overall structure of the graph and thus with the full graph clustering if one would perform such. In particular, it preserves such parameters of the random walk on the full graph as diffusion and commute-time distances between subset nodes. Moreover, working with a small target subset reduces greatly the required dimension of Krylov subspace and allows to exploit the approximations of k-means in the regimes when they are most robust and efficient.

There are several uses for our algorithms. First, they can be employed on their own to clusterize a representative subset in cases when the full graph clustering is either infeasible of simply not required. Second, they may be used for quality control and filtering of noisy data, i.e., outliers. Third, as they drastically reduce the clustering problem size, they enable the application of more sophisticated and powerful approximations of k-means like those based on semi-definite programming (SDP) instead of the conventional Lloyd's algorithm. Finally, they can be used as building blocks of a divide-and conquer type algorithm for the full graph clustering (in progress).

I'll provide the results of numerical experiments with synthetic data as well as real-world statistical data for companies email connections and for citations in ArXiv repository. Time permitting, I'll discuss preliminary results for ongoing project with financial statistics of the stock market data.

This is joint work with **Vladimir Druskin (Worcester Polytechnic Institute) and Alexander Mamonov (University of Houston)**.

Thursday, October 4, 2018
11:00AM-12:00PM
Stratton Hall 203