Regularity Based Spectral Clustering
Advisor: *Prof. Marianna Bolla*

Problem description

Spectral graph theory was developed about 50 years ago (M. Fiedler, D.M. Cvetkovic, F. Chung) to characterize certain structural properties of a graph by means of the eigenvalues of its adjacency or Laplacian matrix. Later on, the eigenvectors corresponding to some near zero eigenvalues of the Laplacian matrix were also used for clustering the vertices into disjoint parts so that the inter-cluster relations are negligible compared to the intra-cluster ones (usual purpose of cluster analysis as a machine learning technique). Since then, the problem was generalized in several ways: to edge-weighted graphs and rectangular arrays of nonnegative entries (e.g., microarrays in biological genetics), to degree-corrected adjacency and Laplacian matrices. Around the millennium, physicists and social scientists introduced modularity matrices and investigated so-called anti-community structures (intra-cluster relations are negligible compared to the inter-cluster ones) in contrast to the former community structures. Going forward, so-called regular cluster pairs with small discrepancy can be defined, where homogeneous clusters are looked for (e.g., in microarrays one looks for groups of genes that similarly influence the same groups of conditions). The existence of such a regular structure is theoretically guaranteed by the Abel-prize winner Szemeredi's regularity lemma that for any small positive ε guarantees a universal number k of clusters (irrespective of the number of vertices) such that partitioning the vertices into k parts (plus possibly a small exceptional one), the pairs have discrepancy less than ε. However, this k can be enormously large and not applicable to practical purposes. Our purpose is to give a moderate k where the discrepancy of the k-clustering dramatically decreases compared to the $k-1$ one. For this, theoretical spectral estimates are at our disposal; see, e.g., *Bolla, M., Spectral clustering and Biclustering, Wiley, 2013* and papers therein on https://www.math.bme.hu/∼marib.

The task is to elaborate the methodology of discrepancy based spectral clustering in the general framework of edge-weighted graphs and rectangular arrays of nonnegative entries. This needs application of existing spectral estimates to assign starting parameters (e.g., number of clusters) to the algorithms. The theory can as well be developed with further estimates. Real life applications are also encouraged (on benchmark datasets or microarrays). In case of success, the findings (theoretical, algorithmic, applications) will be submitted to the special issue Contemporary Spectral Graph Theory of the journal Special Matrices for publication (deadline: 30th November, 2021).

The applicants must have some maturity in graph theory and matrix analysis. The basics will be given in the first lessons. Applicants having good programming skills (e.g. in Python) and ready to implement matrix spectral decomposition and extended k-means clustering algorithms, are also welcome. Tasks are usually shared between the students, but all of them contributing to the joint paper will be coauthors if the paper is accepted.