Impact of Regularization on Spectral clustering
Antony Joseph and Bin Yu
Department of Statistics, University of California, Berkeley

Clustering in networks/graphs is an important problem with applications in the analysis of gene-gene interactions, social networks, text mining, to name a few. Spectral clustering is one of the more popular techniques for such purposes, chiefly due to its computational advantage and generality of application. The algorithm’s generality arises from the fact that it is not tied to any modeling assumptions on the data, but is rooted in intuitive measures of community structure such as sparsest cut based measures (Hagen and Kahng (1992), Shi and Malik (2000), Ng. et. al (2002)).

Here, we attempt to understand regularized form of spectral clustering. Our motivation for this work was empirical results in Amini et. al (2013) that showed that the performance of spectral clustering can greatly be improved via regularization. Here regularization entails adding a constant matrix to the adjacency matrix and calculating the corresponding Laplacian matrix. The value of the constant is called the regularization parameter.

Our analysis is carried out under the stochastic block model (SBM) framework. Under the (SBM) (and its extensions). Previous results on spectral clustering (McSherry (2001), Dasgupta et. al. (2004), Rohe et. al (2011)) also assumed the SBM and relied on the minimum degree of the graph being sufficiently large to prove its good performance. By analyzing the spectrum of the Laplacian of an SBM as a function of the regularization parameter, we provide bounds for the perturbation of the regularized eigenvectors, which, in some situations, does not depend on the minimum degree. For example, in the two block SBM, our bounds depend inversely on the maximum degree, as opposed to the minimum degree.

More importantly, we show the usefulness of regularization in the important practical situation where not all nodes can be clustered accurately. In such situations, in the absence of regularization, the top eigenvectors need not discriminate between the nodes which do belong to well-defined clusters. With a proper choice of regularization parameter, we demonstrate that top eigenvectors indeed discriminate between the well-defined clusters.

A crucial ingredient in the above is the analysis of the spectrum of the Laplacian as a function of the regularization parameter. Assuming that there are $K$ clusters, an adequate gap between the top $K$ eigenvalues and the remaining eigenvalues, ensures that these clusters can be estimated well. Such
a gap is commonly referred to as the eigen gap. In the situation considered in above paragraph, an adequate eigen gap may not exist for the unregularized Laplacian. We show that regularization works by creating a gap, allowing us to recover the clusters.

As an important application of our bounds, we propose a data-driven technique DK-est (standing for estimated Davis-Kahn bounds) for choosing the regularization parameter. DK-est is shown to perform very well for simulated and real data sets.